IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment

BOOK OF SHORT PAPERS

Editors: Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci

Book of Short papers
10th International Conference **IES 2022**
Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment

Department of Economics, University of Campania "L. Vanvitelli",
January 27th - 28th 2022

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

Pietro Amenta -University of Sannio
Matilde Bini -European University of Roma
Luigi D'Ambra -University of Naples "Federico II"
Maurizio Carpita -University of Brescia
Paolo Mariani -University of Milan "Bicocca"
Marica Manisera -University of Brescia
Monica Palma -University of Salento
Pasquale Sarnacchiaro -University of Rome Unitelma Sapienza

Program Committee of the conference IES 2022

Chair: Rosaria Lombardo -University of Campania "L. Vanvitelli"
Fabio Bacchini -ISTAT
Laura Baraldi -University of Campania "L. Vanvitelli"
Eric Beh -University of Newcastle, Australia
Wicher Bergsma -The London School of Economic and Political Science, UK
Enrico Bonetti -University of Campania "L. Vanvitelli"
Eugenio Brentari -University of Brescia
Clelia Buccico -University of Campania "L. Vanvitelli"
Rosalia Castellano -University of Naples "Parthenope"
Ida Camminatiello -University of Campania "L. Vanvitelli"
Carlo Cavicchia -University of Rotterdam, The Netherlands
Enrico Ciavolino -University of Salento
Corrado Crocetta -University of Foggia
Claudio Conversano -University of Cagliari
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Antonio D'Ambrosio -University of Naples "Federico II"
Alfonso Iodice D'Enza -University of Naples "Federico II"
Tonio Di Battista -University of Chieti "G. D'Annunzio"
Michele Gallo -University of Naples "L'Orientale"
Francesco Gangi -University of Campania "L. Vanvitelli"
Michele La Rocca -University of Salerno
Amedeo Lepore -University of Campania "L. Vanvitelli"
Riccardo Macchioni -University of Campania "L. Vanvitelli"
Filomena Maggino -University of Rome "La Sapienza"
Angelos Markos -Demokritus University of Thrace, Greece
Lucio Masserini -University of Pisa
Stefania Mignani -University of Bologna
Nicola Moscariello -University of Campania "L. Vanvitelli"
Francesco Palumbo -University of Naples "Federico II"
Alessandra Petrucci -University of Florence
Alessio Pollice -University of Bari

Donato Posa -University of Salento
Luca Secondi -University of Tuscia
Amalia Vanacore -University of Naples "Federico II"
Michel van de Velden -University of Rotterdam, The Netherlands
Rosanna Verde -University of Campania "L. Vanvitelli"
Donatella Vicari -University of Rome "La Sapienza"
Grazia Vicario -Polytechnic University of Turin
Maurizio Vichi -University of Rome "La Sapienza"

Organizing Committee

Ida Camminatiello -University of Campania "L. Vanvitelli"
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Rosaria Lombardo -University of Campania "L. Vanvitelli"
Elvira Romano -University of Campania "L. Vanvitelli"
Luca Rossi -University Cusano
Violetta Simonacci -University of Naples "L'Orientale"

Editors

Rosaria Lombardo - University of Campania “L. Vanvitelli”, Italy
Ida Camminatiello - University of Campania “L. Vanvitelli”, Italy
Violetta Simonacci - University of Naples “L’Orientale”, Italy

PKE - Professional Knowledge Empowerment s.r.l.
Sede legale: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Sede operativa: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Ufficio Di Rappresentanza: Via Giacomo Peroni, 400 - 00131 Roma (RM)
CF / P.I. 03167830920 — www.pke.it; e-mail info@pke.it — Privacy

February 2022 PKE s.r.l.
ISBN 978-88-94593-35-8 on print
ISBN 978-88-94593-36-5 online
All rights reserved.
This work is protected by copyright law.
All rights, in particular those relating to translation, citation, reproduction in any form, to the use of illustrations, tables and the software material accompanying the radio or television broadcast, the analogue or digital recording, to publication and dissemination through the internet are reserved, even in the case of partial use. The reproduction of this work, even if partial or in digital copy, is admitted only and exclusively within the limits of the law and is subject with the authorization of the publisher. Violation of the rules involves the penalties provided for by the law.
PKE Publisher

Preface

This Book of Short Papers includes all peer-reviewed long-abstracts submitted to the IES2022 conference, titled "Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment", held at the University of Campania "L. Vanvitelli" on January 27-28, 2022. IES2022 is the 10th meeting of the biennial international conference proposed by the permanent group Statistics for the Evaluation and Quality in Services (SVQS) of the Italian Statistical Society (SIS). The SVQS group, born in 2004, focuses on national research programs and applied research activities, on statistical methods and methodologies for the evaluation of the quality of services in public and private fields. For further information, please visit https://www.svqs.it/. IES2022 has been sponsored by the Italian Statistical Society (SIS), the European Network for Business and Industrial Statistics (ENBIS), and the International Association for Statistical Computing (IASC). In addition, also the two SIS groups Statistics and Data Science (SDS) and Enhancement of Public Statistics (VSP) actively supported the conference. IES2022 aims at stimulating a scientific debate on the challenges of Society 5.0 with respect to quality assessment. The conference provides an important moment of reflection for the development of new ideas and methodologies by promoting the rethinking of the open issues in service evaluation within the new paradigm of an interconnected cyber-social system. Service quality assessment represents the starting point for the development of effective policies for private and public institutions, which is crucial for the development of society. Big data, heterogeneous multi-layered structure and designs, cutting-edge analytical tools, and advanced data harvesting techniques have become fundamental for research; nonetheless, they require a continuous effort in terms of proper treatment, interpretation, and supervision to ensure the centrality of human and social problems. In this perspective, IES 2022 main goals are:

- to promote and coordinate the statistical and economic methodologies for the evaluation of a human-centered society emphasizing how statistical thinking, design, and analysis may be of use to a Society 5.0;
- to foster advanced methodological research supporting the assessment of the quality of social services;

- to be a platform where the experts of Statistics, Data Mining, Data Science, Machine Learning, and related disciplines meet for analyzing Big Data.

The high turn-out of the conference, with a total of 107 presentations organized in 22 solicited sessions and 11 contributed sessions, two plenary talks, and the participation of over 300 authors, made evident a very alive interest in evaluation topics. Previous IES editions include:

- IES2009 was held at the University of Brescia (June 24-26, 2009) with selected papers published in special issues of Electronic Journal of Applied Statistical Analysis (EJASA) and Statistica & Applicazioni;
- IES2011 was held at the University of Florence (May 30 – June 1, 2011) with selected papers published in a special issue of the Journal of Applied Quantitative Methods;
- IES2013 held at the University of Milan "Bicocca" (December 9 – 13, 2013) with selected papers published in the Procedia Economics & Finance (Elsevier Publisher);
- IES2015 was held at the University of Bari "Aldo Moro" (June 8 – 9, 2015) with selected papers published in a special issue of Quality & Quantity;
- IES2017 held at the University of Naples "Federico II" (September 6 – 7, 2017) with selected papers published in special issues of Social Indicator Research, Quality & Quantity, and EJASA;
- IES2019 was held at the European University of Rome (July 4 – 5, 2019) with selected papers published in special issues of Socio-Economic Planning Science and EJASA.

All IES2022 contributions are based on the development of innovative statistical methodologies or interesting applications. The topics covered in the numerous presentations range over the following fields: Sustainability, Health, Wellness, Sport, Tourism, Education, Training and Research, Bank and FinTech, Transportation, Environment, Enterprise, Cultural changes and values, Industry and Finance, E-commerce, Digital Marketing, Labour Market, Public Administration, Advertising, Political preferences, Justice System. Several short papers deal with the shock of the COVID-19 pandemic and its impact in different areas such as poverty and sustainability, education and distance learning, student satisfaction, environment, health services, and social interactions. From a methodological standpoint, many of the short papers deal with challenging structures such as high-dimensional data, complex survey designs, constrained variability, sparsity, multicollinearity, and multidimensional longitudinal series. A wide range of statistical tools and models have been employed, including functional data analysis, various types of regression models (high-dimensional, logit, quantile, OLR, LASSO, etc.), machine learning algorithms for classification, methods for multi-way data and contingency tables,

generalized discriminant analysis, multidimensional Item Response Theory, PLS-SEM, advanced visualization techniques, compositional data analysis, Bayesian methods and so on. Extended versions of selected IES2022 papers will be included in a special issue of the Computational Statistics Journal titled "High-dimensional Data Analysis and Visualisation to Assess Service Quality" and of Annals of Operations Research Journal, titled "Statistical Methods and Data-Driven for Decision Making in Public Sector".

Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci
Editors

Luigi D'Ambra
Onorary Coordinator of SVQS Group
Maurizio Carpita
Coordinator of SVQS Group

Contents

Solicited Session SS1 – *Compositional Data Analysis*
Organizer and Chair: Gianna Monti **1**
Rieser C. and Filzmoser P.. *Compositional Data and graph theory* 2
Egozcue J. J., Pawlowsky-Glahn V. and Buccianti A. *Compositional deviations from linear and non-linear equilibria* 8
Monti G. S. and Filzmoser P. *The knockoff filter for FDR control in robust ZeroSum regression in microbiome analysis* 14

Solicited Session SS2 - *Monitoring progress towards SDGs: statistical approaches and methods for measuring poverty, inequalities and food insecurity*
Organizer and Chair: Luca Secondi **18**
Marchetti S., Giusti C., Pratesi M. and Biggeri L. *Poverty indicators adjusted using local price indexes* 19
Tonutti G., Bertarelli G., Giusti C. and Pratesi M. *Assessing the targeting of the anti-poverty measure "Reddito di Cittadinanza" using Small Area Estimation methods* 25
Vargas-Lòpez A. and Secondi L. *Household consumption and food insecurity in Mexico: COVID19 and sustainable development* 31

Session of free contributes SCL1 – *Big Data, Proximity data, Multi way data*
Chair: Donatella Vicari **37**
Metulini R. and Carpita M. *Forecasting Traffic Flows with Complex Seasonality using Mobile Phone Data* 38
Simonacci V., Menini T. and Gallo M. *CP decomposition of 4th-order tensors of compositions* 44
Bove G. *A strategy of analysis of symmetry and skew-symmetry in asymmetric relationships* 50

Session of free contributes SCL2 –*Industry and Society*

Chair: Germana Scepi **56**

Crosato L., Domenech J. and Liberati C. *Toward an early detection of SME's default with websites' indicators* 57

Angelone R. *Italians' Culture and Values after two years of pandemic* . . 63

Aria M., Cuccurullo C., D'Aniello L. and Spano M. *Thematic evolution of Academic Medical Centers' research: a focus on Italian public owned AOUs in metropolitan areas* 67

Session of free contributes SCL3 – *Society and Tourism*
Chair: Emma Zavarrone **73**

Rondinelli R., Palazzo L. and Ievoli R. *Local clustering coefficient to measure intra-regional tourism in Italy* 74

Zavarrone E., M. and Forciniti A. *Local clustering coefficient to measure intra-regional tourism in Italy* 80

Firza N. *Sustainable tourism: The case of Albania* 86

Romano M., Zammarchi G. and Conversano C. *Threshold-based Naïve Bayes Classifier Customer Satisfaction evaluation* 90

Session of free contributes SCL4 –*Society and Innovation*
Chair: Michelangelo Misuraca **95**

Seri E., Alaimo L. S., di Bella E., Cataldo R. and Piscitelli A. *Migrant Integration Policy Index (MIPEX): an analysis of countries via Gaussian mixture modelbased clustering* 96

Rossi L. and Daddi S. *The measure of the BES: a proposal for the aggregation of the indicator education and training* 102

Marino M., Mazza R., Misuraca M. and Stavolo A. *Monitoring consumer sentiment using control charts* 108

Session of solicited contributes SS3 – *Statistical methods for the assessment of student careers in higher education*
Organizer and Chair: Maria Prosperina Vitale **114**

La Rocca M., Niglio M. and Restaino M. *Predicting university students' churn risk* 115

Primerano I., Santelli F. and Usala C. *Discovering archetypal universities in Italian higher education mobility flows* 121

Cascella C. and Ragozzini G. *Measuring quality of students' careers in Higher Education: a systematic literature review* 127

Porcu M., Sulis I. and Usala C. *Estimating the peers effect on students' university choices* 134

Session of solicited contributes SS4 – *Conformity assessment and quality predictions- itENBIS*
Organizer and Chair: Amalia Vanacore **140**
Vanacore A., Pellegrino M. S. and Ciardiello A. *Testing the predictive performance of multi-class classifiers* 141
Borgoni R., Gilardi A. and Zappa D. *Optimal Subgrids from Spatial Monitoring Networks* 148
Pennecchi F. and Kuselman I. *Extension of the JCGM 106:2012 - Conformity assessment of multicomponent items and finite statistical samples* 153

Session of solicited contributes SS5 – *Big Data and Large-dimensional Data*
Organizer and Chair: Stefania Mignani **160**
Farné M. and Vouldis A. *ROBOUT: a conditional outlier detection methodology for large-dimensional data* 161
Camillo F. *Behaviours, emotions and opinions in modern citizen or customer relationship systems: a correct integration of small and big data for hyper-targeting, personal advertising and look-alike* 168
Camminatiello I. and Lucadamo A. *A model for assessing sea environmental quality* 172

Session of solicited contributes SS6 – *Health Quality*
Organizer and Chair: Paolo Mariani **178**
Bartolini B., Bertoldi S., Benedan L., Galeone C., Mariani P., Sofia F. and Zenga M. *The uneasiness index in a patient-designed quality of life questionnaire* 179
Benedan L., Hachem M. E., Galeone C., Mariani P., Pilo C. and Tadini G. *Assessing the Quality of Life of patients with Epidermolysis Bullosa (EB): Development of a patient-centered questionnaire* 184
Marletta A. and Morandi M. *Survival analysis in a business context: how to control the abandons of my subscribers* 190

Session of free contributes SCL5 – *Assessing Performance*
Chair: Cristina Davino **194**
di Trapani G. *Political performance measuring and tracking through a system based on the Political Performance Indicator (Iep): Naples 2021 case* 195
Montanari G. E. and Doretti M. *A class of case-mix adjusted probability-based indices for performance evaluation* 202

Cavicchia C., Sarnacchiaro P., Vichi M. and Zaccaria G. *An ultrametric model to build a Composite Indicators system* 208

Session of free contributes SCL6 – *Statistical Learning*
Chair: Massimo Aria **212**

Migliorati M. and Brentari E. *Feature definition for NBA result prediction through Deep Learning* 213

Levantesi S., Lizzi M. and Nigri A. *An application of contrast trees for mortality models diagnostic and boosting* 219

Aria M., Gnasso A. and D'Aniello L. *Twenty Years of Random Forest: preliminary results of systematic literature review* 225

Park H., Hong J., Shin Y. and Park J-S. *A study on the GEV activation function for classification of class imbalance data* 231

Session of free contributes SCL7 – *Society and Disparity*
Chair: Pasquale Sarnacchiaro **236**

Alaimo L. S., D'Urso P. and Nigri A. *The gender gap in lifespan disparity as a social indicator of international countries: A fuzzy clustering analysis approach* 237

Iannario M. and Tarantola C. *Modelling scale effects via a Bayesian approach: an application to decision making in public sector* 243

Gangi F. , Daniele L. M. and Coscia M. *Board Gender Diversity and Social engagement: evidence from the banking industry* 249

Session of free contributes SCL8 – *Education*
Chair: Matilde Bini **256**

Davino C. and Lamberti G. *Assessing heterogeneity in students' performance. The case of the Massive Open Online Courses* 257

Primerano I., Catone M. C., Giordano G. and Vitale M. P. *Assessing undergraduate students' perceptions of distance learning during the COVID-19 pandemic* 263

Crisci A., Lucadamo A. and Amenta P. *PhD satisfaction analysis in Italian University via Classification tree, Bagging and Random Forest* 269

Cervellera S., Cusatelli C. and Giacalone M. *Comparative Analysis of Student Learning: Technical, Methodological and Result Assessing of PISA-OECD and INVALSI-Italian Systems* 275

Session of solicited contributes SS7 – *SEM with PLS: Theory and Applications*
Organizer and Chair: Enrico Ciavolino **281**

Wang S., Cheah J. and Roldan J. L. *PLS-SEM basics and its potential applications: A quick journey* 282
Cefis M. and Carpita M. *A PLS-SEM confirmatory composite analysis for football goalkeeper's performance validation* 288
Pasca P., Misuraca M., Meloni A. and Ciavolino E. *Text-mining and PLS-SEM combination to measure food satisfaction with Google Review: When the gut (re)counts!* 294

Session of solicited contributes SS8 – *Applications of non standard statistical tools to real-life*
Organizer and Chair: Antonio D'Ambrosio **300**
Iorio C. and Pandolfo G. *A robust strategy for building a financial portfolio* 301
Ruscone M. N. and De Luca G. *Conditional copula a financial application* 307
Ortu M., Frigau L. and Contu G. *Explaining Student Satisfaction Assessments: A Natural Language Processing Approach* 313

Session of solicited contributes SS9 – *Innovation and Value Co-creation in Society*
Organizers and Chairs: Alessandra De Chiara & Anna D'Auria **319**
Mauro S. *The smart working towards a Society 5.0* 320
Del Vacchio E., Carignani F., Laddaga C. and Bifulco F. *Innovative interaction in Society 5.0: insights from the cultural sector* 326
D'Auria A. and De Chiara A. *Society 5.0: a bibliometric analysis* 333

Session of solicited contributes SS10 – *Statistical learning for mainstream press, health and fiscal data*
Organizer and Chair: Claudio Conversano **339**
Baldassarre A. and Carullo D. *The regression trunk model for partitioning Italian municipalities based on their fiscal capacities and its determinants* 340
Pandolfo G. and Iorio C. *An analysis of Italian healthcare mobility through a depth-based clustering procedure* 346
Dossou B. F. P. and Wilhelm A. F. X. *Automatic Fake News Detection to Ensure Quality of News Articles* 352

Session of solicited contributes SS11 – *Multi-way Methods for Evaluation Service*
Organizer and Chair: Michele Gallo **358**
Simonacci V., Marino M., Grassia M. G. and Gallo M. *Multiple factor analysis with external information on PISA survey data* 359

Bocci L. and D. Vicari D. *A three-way analysis of well-being in Italy over time* 365
Cerqueti R., Mattera R. and Scepi G. *Multiway approach for clustering time series with time varying parameters* 371

Session of solicited contributes SS12 – *Assessment of Management Quality*
Organizer and Chair: Clelia Fiondella **377**
Belfiore A., Cuccurullo C. and Aria M. *Using partial triadic analysis for depicting the temporal evolution of Italian private healthcare organizations* 378
Bollani L., Celegato A., Barbero F. and Fontemaggi F. *Management of the human factor into the company. An experience from the aeronautic sector.* 385

Session of solicited contributes SS13 – *New technologies for students learning assessment and evaluation*
Organizer and Chair: Alfonso Iodice D'Enza **391**
Themelis E. and Markos A.*A non parametric cognitive diagnostic method in classroom assessment conditions* 392
Iannario M., Iodice D'Enza A. and Romano R. *Hybrid unfolding models to Likert-scale data to assess distance learning perception in higher education* 398
Pacella D., Fabbricatore R., Galluccio C. and Palumbo F. *Classification of Statistics learners using multi-dimensional latent class IRT model and archetypal analysis: the ALEAS app* 404

Session of solicited contributes SS14 – *Labor Market and Enterprises*
Organizer and Chair: Lucio Masserini **408**
Maggioni G., Mariani P., Marletta A. and Zenga M. *Searching for new trends and dynamics in Labour Market: a statistical approach for the recruiting process* 409
Bruttini P., Mariani P., Marletta A., Masserini L. and Zenga M. *A new definition of the professional figure Open Manager* 413

Session of solicited contributes SS15 – *Statistical Approaches to Environmental Sustainability*
Organizer and Chair: Alfonso Piscitelli **417**
Aspinall R. *Measuring sustainability as an emergent property of whole system dynamics* 418

Alaimo L. S. and Finocchiaro G. *Tourism sustainability in the Italian regions: a fuzzy approach* 424
D'Uggento A. M. *How young people perceive environmental issues, react to ecological concerns and commit themselves to sustainable behaviours* 430

Session of solicited contributes SS16 – *Statistical Methods for Environmental, Natural Resources and Health Assessment*
Organizer and Chair: Alessio Pollice **437**
Ferretti A., Ippoliti L. and Valentini P. *Spatial-ARFIMA models for the statistical analysis of environmental lattice processes* 438
Arima S., Pasculli G. and Polettini S. *A Bayesian non parametric approach for bias correction for underreported data* 444
Lasinio G. J., Mastrantonio G., Pollice A., Ventura D., Mancini G. and Ardizzone G. *Assessment of the impact of anthropic pressures on the Giglio island meadow of Posidonia oceanica* 450

Session of solicited contributes SS17 – *Functional Data Analysis Methodologies for Quality Assessment*
Organizer and Chair: Elvira Romano **456**
Fortuna F., Maturo F. and Di Battista T. *Improving the quality of questionnaires via the combined use of functional outlier detection and Item Response Theory* 457
Naccarato A., Fortuna F. and Terzi S. *Assessing government effectiveness over time: a functional data analysis approach* 462
Balzanella A. and Verde R. *Mining Distributed Acoustic Sensing data for vehicle traffic monitoring* 467

Session of solicited contributes SS18 – *Evaluation and assessment of cognitive and learning processes*
Organizer and Chair: Francesco Palumbo **473**
Ponticorvo M. and Argiuolo T. *Neuropsychological Assessment supported by Technology: the E-BTT case* 474
Milano N. and Gigliotta O. *Mining Introducing OpenAi-ES in Interactive Data. Clustering with R-EVOK* 480
Davino C., Gherghi M., Palumbo F. and Vistocco D. *Modeling heterogeneity in student's satisfaction during the Covid-19 pandemia* . . . 484

Session of solicited contributes SS19 – *Substainability and Environment*
Organizer and Chair: Ida Camminatiello **490**

Lombardo R. and Beh E. J. *Partitioning the Cressie-Read divergence statistic for three-way contingency tables: a study on environmental sustainability data* 491
Prahadchai T., Hong J., Busababodhin P. and Park J-S.. *Analysis of maximum precipitation in Thailand using non-stationary extreme value models* 498
Tregua M. and Scaglione M. *Assessing citizens' participation to urban transformation: a review of quantitative methods* 504

Session of solicited contributes SS20 – *Statistical methods for health and environmental impact assessment*
Organizer and Chair: Fabrizio Maturo **512**
Gattone S. A. and Di Battista T. *Density estimation via Functional Data Analysis* 513
Evangelista A., Acal C., Aguilera A. M., Sarra A., Di Battista T. and Palermi S. *A new multivariate functional ANOVA approach for assessing air quality data amid COVID-19 pandemic* 517
Diana A., Romano E. and Irpino A. *Conformal Prediction for Geographically Weighted Functional Regression models: an application for environmental impact assessment.* 523

Session of solicited contributes SS21 – *Local sustainability assessment: challenges in building quality indicators*
Organizers and Chairs: Francesca Fortuna & Alessia Naccarato **529**
Grimaccia E. *Urban Sustainability Assessment: A Proposal for an Index Based on SDGs' Indicators* 530
Liberati P. and Resce G. *Between and Within Country Inequality in Regional Well Being* 536
Di Battista T., Nissi E. and Sarra A. *Equitable and sustainable well-being over time: a functional approach* 542

Session of free contributes SCL9 – *Health and Covid-19*
Chair: Maria Sole Pellegrino **548**
Cao N., Calcagní A. and Finos L. *Twitter about COVID-19: An application of Structural Topic Models to a sample of Italian tweets* 549
Parretti C., Tartaglia R., Sbrana G., Mandó M. and Pacchi S. *Assessing the quality of a health service through the risk profile number (RPN)* 555
Di Lorenzo G., Franchetti G. and Politano M. *The insurance premium structure for a covid-19 insurance policy* 562

Session of solicited contributes SS22 – *Statistics, culture and tourism*
Organizer and Chair: Marica Manisera **568**
Cristiani P. *How data can influence the promotion and the consumption of cultural experience* 569
Carpita M., Manisera M. and Zuccolotto P. *Mobile phone data to monitor the impact of social and cultural events of Brescia* 575
Capecchi S., Quaranta G. and Salvia R. *Analysing opinions on sustainable tourism in the Vallo di Diano area, Campania, Italy* 582

Session of free contributes SCL10– *Time Series data, Panel data and Circular Economy*
Chair: Anna Crisci **588**
Scaccabarozzi D., Toninelli D., Zurlo D., Bacchini F. and Iannaccone R. *Testing the Participation Gap Inclusion within the Wage Phillips Curve* 589
Fusco D., Liguori M. A. and Moretti V. *Multisource approach for trends evaluation. An application at the agricultural sector* 596
Bonnini S. and Borghesi M. *A permutation test on the relationship between Circular Economy and firm size* 601
Mele S., Izzo F. and Tomnyuk V. *Circular economy and business models: a literature review* 607

Session of free contributes SCL11 –*Modelling Extreme Values, High dimensional, time series data*
Chair: Violetta Simonacci **611**
Shin Y., Busababodhin P. and Park J-S. *Modeling extreme values using the r-largest four parameter distribution* 612
Sabri M., Maturo F., Verde R., Riffi J. and Yahyaouy A. E. *Classification of ECG signals based on functional data analysis and machine learning techniques* 618
Park J-S., Shin Y., Shin Y. and Hong J. *Determining shape parameters in a climate multi-model ensemble* 624

Solicited Session SS1 – *Compositional Data Analysis*
Organizer and Chair: Gianna Monti

Compositional Data and graph theory

Dati Compositivi e Teoria dei grafi

Christopher Rieser and Peter Filzmoser

Abstract In this short paper we discuss an extension of compositional data to signals with network domain. We recapture the geometric nature of compositional data and describe its relationship to graphs. The derived methodology is illustrated with a data set originating from the Gemas project. This data set with concentrations of chemical elements in soil samples has been considered multiple times in the literature, and we present new insights by using this connection of compositional data analysis with graph theory.

Key words: Compositional Data, Graph theory

1 Introduction

Compositional data analysis (CoDa) has been a very active field of research since the original work of John Aitchison [1]. Assume that $x = (x_1, \ldots, x_D)'$ is a D-dimensional multivariate strictly positive variable of interest, then the core assumption in CoDa is that the information we are interested in is carried by all pairwise log-ratios $\log(\frac{x_i}{x_j})$, for $i, j = 1, ..., D$. This point of view led to the development of the Aitchison geometry and the adaption of tools from classical multivariate statistics to the compositional framework. Many data sets, such as Microbiome data [4] or chemical compositions, have been recognized to bear a compositional nature and have to be treated accordingly. In reality, however, the assumption that all pairwise log-ratios are equally important and influential in the analysis does not seem to be

Christopher Rieser
Institute of Statistics & Mathematical Methods in Economics
Vienna University of Technology, e-mail: christopher.rieser@tuwien.ac.at

Peter Filzmoser
Institute of Statistics & Mathematical Methods in Economics
Vienna University of Technology e-mail: peter.filzmoser@tuwien.ac.at

appropriate. In practice it seems more realistic that only specific log-ratios are relevant for the analysis, and some log-ratios shall not even be considered because the corresponding compositional parts might not present any interpretable relationship. This naturally leads to a consideration in form of a graph structure of the relevant connections between compositional parts, and an approach with links graph theory and CoDa has been proposed in [10]. In this paper we use this approach and present an application to a well studied compositional data set from the Gemas project, a European geochemical mapping project, where the chemical element concentrations of more than 2000 soil samples have been analyzed. The data set is freely available in the R package [11].

2 Some important concepts from CoDa

We denote $\mathbb{R}^D_+$ as the space of strictly positive D-dimensional real valued vectors. In classical CoDa one works in the D-part simplex $\mathscr{S}^D$

$$\mathscr{S}^D := \left\{ (x_1,\dots,x_D)' \in \mathbb{R}^D_+ \middle| \sum_{j=1}^{D} x_j = 1 \right\} \subset \mathbb{R}^D_+ ,$$

equipped with the two operations $x \oplus y := (x_1 y_1,\dots,x_D y_D)'$ and $\alpha \odot x := (x_1^\alpha,\dots,x_D^\alpha)'$, for any $x = (x_1,\dots,x_D)'$, $y = (y_1,\dots,y_D)' \in \mathbb{R}^D_+$ and $\alpha \in \mathbb{R}$, and the Aitchison inner product,

$$\langle x,y \rangle_{\mathscr{A}} := \frac{1}{2D} \sum_{i,j=1}^{D} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right) . \tag{1}$$

The inner product (1) being at the core of the Aitchison geometry $(\mathscr{S}^D, \langle\cdot,\cdot\rangle_{\mathscr{A}}, \oplus, \odot)$ has the important property of scale invariance – any rescaling of x or y by a constant will not change the analysis. Further desirable properties are permutation invariance and subcompositional coherence [1]. The space $(\mathscr{S}^D, \langle\cdot,\cdot\rangle_{\mathscr{A}}, \oplus, \odot)$ can be shown to be a Hilbert space as well as one-to-one isometrically to $(\mathbb{R}^{D-1}, \langle\cdot,\cdot\rangle_E, +, \cdot)$, where $\langle\cdot,\cdot\rangle_E$ is the standard Euclidean Inner product. Multiple isometries exist, often considered are so called ilr (isometric logratio)-maps, see [2], which are given after fixing a matrix $\mathbf{V} \in \mathbb{R}^{D\times D-1}$ with orthogonal columns spanning the space $\{z \in \mathbb{R}^D | \langle z,1\rangle_E = 0,\}$, by

$$\mathrm{ilr}_{\mathbf{V}} : \mathscr{S}^D \to \mathbb{R}^{D-1}, \quad \mathrm{ilr}_{\mathbf{V}}(x) := \mathbf{V}' \,\mathrm{clr}(x) , \tag{2}$$

where clr denotes the centered log-ratio map

$$\mathrm{clr} : \mathscr{S}^D \to \mathbb{R}^D, \quad \mathrm{clr}\,(x) := \left(\log\left(\frac{x_1}{\sqrt[D]{\prod_{j=1}^D x_j}}\right), ..., \log\left(\frac{x_D}{\sqrt[D]{\prod_{j=1}^D x_j}}\right)\right)' . \tag{3}$$

Being isometries, any ilr map has the following properties

$$\text{ilr}(x \oplus y) = \text{ilr}(x) + \text{ilr}(y) \tag{4}$$
$$\text{ilr}(\alpha \odot x) = \alpha \, \text{ilr}(x) \tag{5}$$
$$\langle x, y \rangle_{\mathscr{A}} = \langle \text{ilr}(x), \text{ilr}(y) \rangle_E \tag{6}$$

for any $x = (x_1, \ldots, x_D)', y = (y_1, \ldots, y_D)' \in \mathbb{R}^D_+$ and $\alpha \in \mathbb{R}$. For a more detailed introduction to CoDa we refer to [3].

The Aitchison inner product (1) contains the information of all log-ratios which is in certain settings undesirable. Generalizing (1) leads to considering instead an inner product that only contains the information of certain pairs of ratios. Such an extension can be made by means of graph theory by defining an inner product on a subspace of the strictly positive variables as

$$\langle x, y \rangle_{\mathbf{W}} := \log(x)' \mathbf{L}_{\mathbf{W}} \log(y),$$

see [10], where $\mathbf{L}_{\mathbf{W}}$ is the Laplacian matrix to a graph induced by weights $\mathbf{W} = (w_{ij})_{1 \le i,j \le D}$, see [7]. The entries of the matrix $\mathbf{W}$ correspond to the weights one wants to put on different log-ratios. Similar to the compositional case this approach allows for the construction of isometric maps.

2.1 Analysis of data from the Gemas project

We consider in the following a data set from the Gemas project, see [8] and [9], which is freely available in the R package robCompositions [11]. It contains information about 2108 soil samples taken at various locations throughout Europe. Each soil sample has been analyzed for the chemical composition and the concentration of the 18 elements Al, Ba, Ca, Cr, Fe , K, Mg, Mn, Na, Nb, P, Si, Sr, Ti, V, Y, Zn, Zr, in in mg/kg, has been recorded together with the sample location.

Figure 1 shows the graph of all elements as nodes with the log-ratios that have been included in the analysis as edges. These log-ratios were chosen such that high variation between log-ratios in the data gets weighted highly while retaining the information by all pairwise log-ratios. This is similar to a stepwise approach considered in [6]. Some elements such as Calcium have seemingly a more central role with multiple edges to other elements. Since the line thickness reflects the size of the weight, Figure 1 also displays that high weight is put, for example, on the ratios (Calcium, Zirconium), (Calcium, Silicon) and (Magnesium, Zirconium), due to a high variation of the log-ratio between these elements in the data. Given a graph structure and its weighs we can construct an isometric map to the Euclidean space in a similar way as in (2).

Figure 2 shows the first three coordinates of this Graph-ilr map in addition to the explained variance and the procrustes score, see [5], relative to the original data set. Each coordinate is of the form $\log(\prod_{i=1}^D x_i^{\rho_i})$, where ρ_i is plotted on the upper left

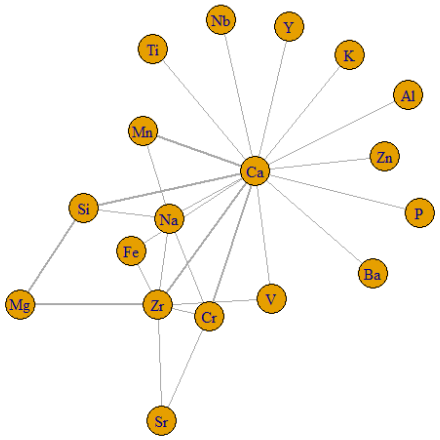


Fig. 1 Graph of chemical elements. Edges indicate which log-ratios have been used for the analysis. Thicker edges correspond to higher weights.

corner of the plots. For better readability we have scaled the values to lie in $[-1, 1]$. We see that the first coordinate, accounting for 39% of the variance, weights Calcium very highly. Compared to all the other elements, Calcium is prevalent in the East and South-East of Spain as well as in the South of France, and in some parts of Italy and Greece. In the Northern countries there are no elevated concentrations. The second coordinate, accounting for an additional 14% of variance, is heavily influenced by the magnitude of Zirconium, and to some extent Sodium as well as Magnesium. It is, comparatively to the other elements, low in the Eastern and Central parts of Central Europe and higher in Scandinavia and Southeast as well as in Southern Europe. The third coordinate accounts for another 18% of explained variance. This coordinate is strongly influenced by the magnitude of Sodium and Chromium as well as lightly by Zirconium and Silicon. A higher third coordinate indicates that Sodium and Zirconium are elevated, relative to the weighted geometric average of the other elements. This is the case for Finland, Sweden and Norway, but also the islands of Sardinia and Corsica as well as a small region at the Spanish-Portuguese border. The first three coordinates together explain up to 71% of the variance and have a procrustes score of 84%. As there are fourteen more coordinates, we can say that a lot of the total information of all pairwise log-ratios is retained in these three coordinates already. For comparison, the first three principal components of the classical CoDa approach explain slightly more variance, at 74%, with a procrustes score of 86%.

Acknowledgements This research was supported by the Austrian Science Fund (FWF) under the grant number P 32819 Einzelprojekte.

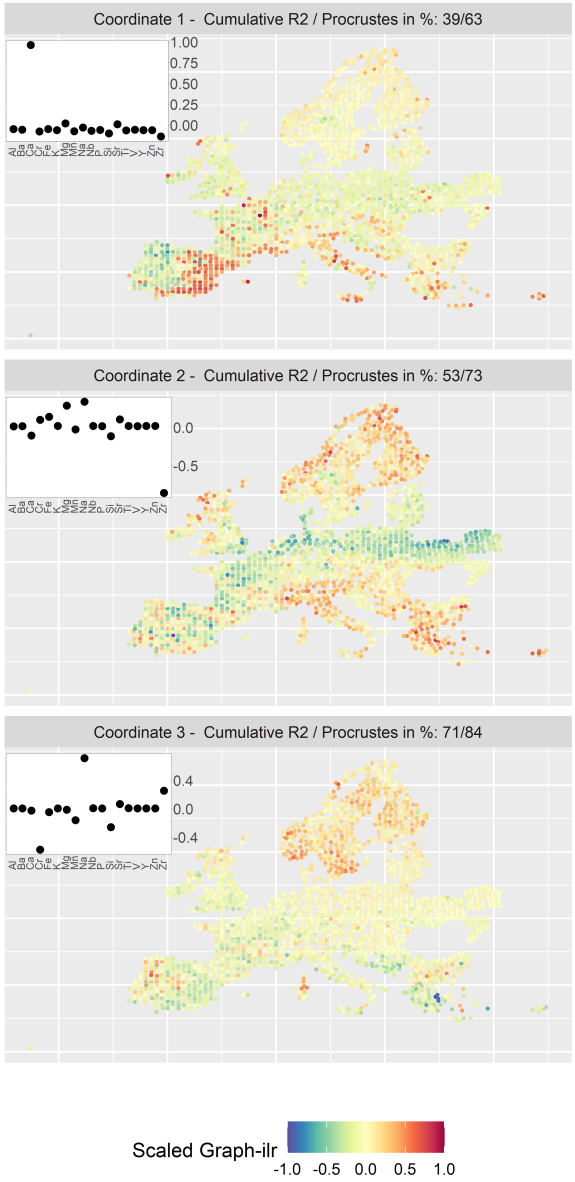


Fig. 2 Scaled values of the first three coordinates of the Graph-ilr map. The cumulative explained variance as well as the procrustes score of the coordinates relative to the original data set is at the top of every plot. On the upper left we plot the weighs of the elements for each coordinate.

3 Conclusions

In this short paper we first started by recalling the most important properties of CoDa and the Aitchison geometry, as well as the graph theoretical extension. We then analyzed a geochemical data set from this new perspective and saw that without any loss of information we could consider this data set as a graph compositional one. The (graphical) ilr coordinates were interpretable and almost as informative as the projection onto the first three principal component loadings of classical CoDa.

References

1. Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) **44**(2), 139–160 (1982)
2. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C.: Isometric logratio transformations for compositional data analysis. Mathematical Geology **35**(3), 279–300 (2003)
3. Filzmoser, P., Hron, K., Templ, M.: Applied compositional data analysis. Springer (2018)
4. Gloor, G.B., MacKlaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: and this is not optional. Frontiers in Microbiology **8**, 2224 (2017)
5. Gower, J.C., Dijksterhuis, G.B., et al.: Procrustes problems, vol. 30. Oxford University Press (2004)
6. Greenacre, M.: Variable selection in compositional data analysis using pairwise logratios. Mathematical Geosciences **51**(5), 649–682 (2019)
7. Mohar, B.: The laplacian spectrum of graphs. In: Graph Theory, Combinatorics, and Applications, pp. 871–898. Wiley (1991)
8. Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P.: Chemistry of Europe's agricultural soils, Part A: Methodology and interpretation of the gemas data set. Geologisches Jahrbuch (Reihe B) **102**, 523 (2014)
9. Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P.: Chemistry of Europe's agricultural soils–Part B: General background information and further analysis of the GEMAS data set. Geologisches Jahrbuch (Reihe B) **103**, 352 (2014)
10. Rieser, C., Filzmoser, P.: Graphs and compositional data. Tech. Rep. CS2021-10/1, TU Wien, Institute of Statistics and Mathematical Methods in Economics (2021)
11. Templ, M., Hron, K., Filzmoser, P.: robCompositions: an r-package for robust statistical analysis of compositional data. In: Compositional Data Analysis: Theory and Applications, pp. 341–355. John Wiley and Sons (2011)

Compositional deviations from linear and non-linear equilibria

Deviazioni composizionali da equilibri lineari e non lineari

J. J. Egozcue, V. Pawlowsky-Glahn and A. Buccianti

Abstract Equilibrium of components is an important issue. Well known examples are the Hardy-Weinberg law in population dynamics and chemical equilibrium regulated by stoichiometry in mineralogy. These equilibria appear as a restriction in the sample space of compositional data. Hardy-Weinberg equilibrium corresponds to a linear equilibrium, while olivine crystals have a stoichiometry that corresponds to a non-linear one. The basic concepts on which our model is based are introduced. The non-linear cases admit a linearisation by adding terms to the original composition. The deviation to the equilibrium locus is shown to be the Aitchison distance between an observation and its orthogonal projection on that locus.

Abstract *Lo studio dell'equilibrio tra componenti differenti di un sistema è un noto campo di ricerca. Esempi sono dati dalla legge di Hardy-Weinberg in dinamica delle popolazioni e dalla stechiometria in mineralogia. Questi equilibri appaiono come una restrizione nello spazio campionario delle composizioni. L'equilibrio di Hardy-Weinberg è rappresentato da una linea retta nel simplex mentre la stechiometria delle olivine mostra un equilibrio non lineare. I concetti base su cui il modello di questi equilibri è basato sono presentati. I casi non lineari ammettono una linearizzazione addizionando alcuni termini alla composizione originale. La deviazione dal luogo dell'equilibrio è mostrata essere la distanza di Aitchison tra una osservazione e la sua proiezione ortogonale da questo sito.*

Keywords: compositional data, Aitchison geometry, sample space

J. J. Egozcue
Technical University of Catalonia, Barcelona (Spain),
e-mail: juan.jose.egozcue@upc.edu

V. Pawlowsky-Glahn
University of Girona, Girona (Spain) e-mail: vera.pawlowsky@udg.edu

A. Buccianti
Universitá degli Studi di Firenze, Firenze (Italy) e-mail: antonella.buccianti@unifi.it

1 Introduction

In [5] the idea of deviation from compositional equilibrium was developed and some techniques of linearisation in the non-linear case were discussed. The key concept for such a development is the sample space and its structure. In the case of compositional data, an appropriate sample space is the simplex endowed with a particular, Euclidean, geometry, termed *Aitchison geometry* [10] because the main elements were introduced by J. Aitchison [1, 2]. Later developments can be found in [11] and [4], as well as in references therein.

Compositions, as elements of a Euclidean space, can be represented by coordinates, specifically by orthonormal Cartesian coordinates. These are called isometric log-ratio or orthonormal, ilr or olr, coordinates. A particular type of olr coordinates, termed *balances*, are obtained applying a Sequential Binary Partition (SBP) [11]. A balance of two groups of parts is a normalised log ratio of geometric means of parts within each group. Once compositions are represented by their olr coordinates, distances and projections are computed as in a standard Euclidean space.

2 Deviations from compositional equilibrium

Equilibrium can have different meanings depending on the context. Here *equilibrium* refers to a restriction in the sample space. In the compositional case, when realisations of a random composition are limited to span a locus within the sample space it is said they are in equilibrium. The subset is the equilibrium locus. The origin of the equilibrium locus can be diverse. It may correspond to physical situations where the samples are attracted or pushed away; or it may be defined subjectively by the analyst. In practice, compositional samples seldom satisfy exactly a well-defined equilibrium locus, and interest is then focused on the deviations from equilibrium.

The simplest equilibrium loci are given by a single point of the sample space. For instance, consider D shares of a commodity. Equality means that all shares are equal, that is, *equality* is the neutral element, the origin of the sample space. Then, interest is on departures from *equality*, taken as equilibrium, and consequently, on *indices of inequality* like the Gini index [7] or, more close to the compositional approach, that defined in [4]. More general cases are those in which the equilibrium locus is a hyperplane of the simplex. The points $\mathbf{x}$ of the D-part simplex in an hyperplane satisfy

$$\sum_{i=1}^{D} \alpha_i \log x_i = K \,, \quad \sum_{i=1}^{D} \alpha_i = 0 \,,$$

where K is some real value. Yet, this approach is too general, since it lacks sparsity and simplicity [9]. As an alternative, attention is centred in hyperplanes defined as a constant balance (see [5] and Section 3). Then, the deviation from the equilibrium locus is

$$\delta_{eq} = \sqrt{\frac{rs}{r+s}}\ \log\frac{\mathrm{g_m}(G)}{\mathrm{g_m}(H)} - K\,,$$

where $\mathrm{g_m}(\cdot)$ denotes the geometric mean of the argument and G and H are groups of parts. Note that δ_{eq} can be considered as a real random variable that allows statistical inference about it when a sample is available (Fig. 1).

There are non-linear cases of equilibrium loci which are both sparse and simple. For example, a constant log ratio which is not scale invariant or which includes amalgamation of parts, thus defining a warped surface. The linearisation technique consists in expanding the original composition with new parts including terms necessary to form the constant log ratio defining the equilibrium locus (see Section 3). Once the equilibrium locus is expressed as a constant balance of the expanded composition, the problem of computing deviations from equilibrium is reduced to the linear case.

3 Examples

(1) *Hardy-Weinberg equilibrium (HWE)* is a well-known genetic law [8, 12]. It states that the proportion of genotypes AA, BB, AB in a closed population under random mating is $(p_A^2,\, p_B^2,\, 2p_Ap_B)$, where p_A, p_B, are the proportions of alleles A, B in the parent population. HWE is then expressed as

$$B(AA, BB/AB) = \sqrt{\frac{2}{3}}\ \log\frac{(p_A^2 p_B^2)^{1/2}}{2p_Ap_B} = \sqrt{\frac{2}{3}}\ \log(1/2) = K\ ,$$

that is the balance $B(AA, BB/AB) = K$ under exact equilibrium. Consequently, $\delta_{eq} = B(AA, BB/AB) - K$ measures the distance or deviation from equilibrium. The set of olr coordinates is completed with $B(AA/BB)$, which can take any real value depending on p_A and p_B.

A data set of frequencies of blood types MM, NN, MN in several populations is used for illustration. The data come from Gower (1987), referenced in [3]. Figure 1 (left) shows the frequencies of blood types in different populations represented by different colours. The line in red is the Hardy-Weinberg equilibrium locus. This data set can be represented using the following olr coordinates

$$b_1 = \sqrt{\frac{1}{2}}\ \log\frac{MM}{NN}\quad,\quad b_2 = \sqrt{\frac{2}{3}}\ \log\frac{\sqrt{MM\cdot NN}}{MN}\ ,$$

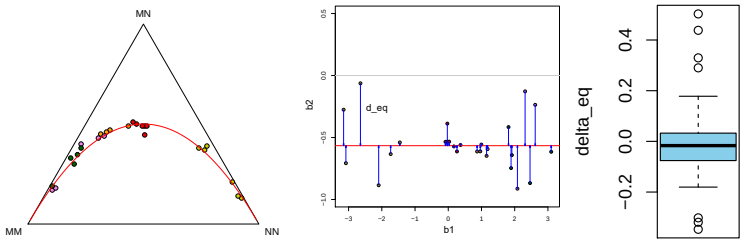


Fig. 1 Left: HWE data, raw proportions. Middle: HWE data in olr coordinates. Right: Box plot of δ_{eq}. See text for details.

where the Hardy-Weinberg equilibrium corresponds to $b_2 = K$ (Fig. 1, middle). The red line is the HWE locus, and the blue arrows are the deviations δ_{eq} from HWE. The boxplot at the right corresponds to the sample of δ_{eq}'s.

(2) *Stoichiometric equilibrium.* Crystals of olivine are formed by atoms of Si combined with 2 atoms of Fe and/or Mg, which can substitute each other. The composition, expressed in proportions of atoms, is here denoted ([Si], [Fe], [Mg]). The stoichiometric equilibrium of Ferric-Magnesic olivines is expressed as

$$B([\mathrm{Si}]/([\mathrm{Fe}]+[\mathrm{Mg}])) = \sqrt{\frac{1}{2}}\ \log\frac{[\mathrm{Si}]}{[\mathrm{Fe}]+[\mathrm{Mg}]} = \sqrt{\frac{1}{2}}\log\frac{1}{2} = K\ . \qquad (1)$$

The equilibrium locus depends on the units of the elements. The value of [Fe]+[Mg] is not proportional to Fe + Mg in ppm, as amalgamation is not a linear function in the simplex [4], and the equilibrium locus defined by Equation (1) is a non-linear manifold in the Aitchison geometry of the simplex.

To define a deviation from the stoichiometric equilibrium using the technique described for linear cases, the term [Fe] + [Mg] can be included as a part in a new composition ([Si], [Fe], [Mg], [Fe] + [Mg]). When representing this new composition in olr coordinates, an appropriate SBP can produce the balance $B([\mathrm{Si}]/([\mathrm{Mg}]+[\mathrm{Fe}]))$. Hence, δ_{eq} is the difference between this balance and the value of K. Note that the Aitchison geometry used is that of the new 4-part composition and not that of the 3-part original one. For illustration a set of 1180 olivine samples in mg/kg [6] has been selected and converted to atoms using the average atomic weight of each element (see [5]).

Figure 2 shows the olivine sample in a ternary diagram following approximately the red line corresponding to the stoichiometric equilibrium. Straight lines in a ternary diagram are curved in a representation in olr Cartesian coordinates. For instance, using

Compositional deviations from linear and non-linear equilibria

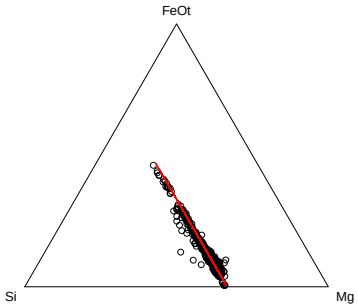


Fig. 2 Olivine data (Mg,Fe,Si) in (atoms) represented in a ternary diagram. The red line is the stoichiometric equilibrium locus.

$$b_1 = \sqrt{\frac{1}{2}}\ \log\frac{[\mathrm{Mg}]}{[\mathrm{Fe}]} \quad , \quad b_2 = \sqrt{\frac{2}{3}}\ \log\frac{[\mathrm{Si}]}{\sqrt{[\mathrm{Mg}]\cdot[\mathrm{Fe}]}},$$

the left panel of Figure 3 is obtained. Now the equilibrium locus (red line) is

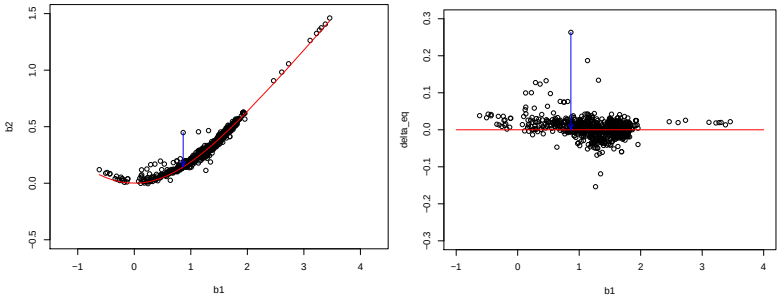


Fig. 3 Olivine data (Mg,Fe,Si) in (atoms) in olr coordinates,left panel. The red line is the stoichiometric equilibrium locus. Right panel: δ_{eq} for each observation ordered by b_1.

curved. The blue arrow shows the projection of one point on the equilibrium line. The length of the arrow is not proportional to δ_{eq}. The values of δ_{eq} are shown in Figure 3 (right panel), where the length of the blue arrow is now proportional to δ_{eq}. The red line represents the equilibrium after linearisation.

4 Conclusions

Compositional equilibrium is defined as a restriction of the sample space. The restriction, or equilibrium locus, can be a very general set, but interest is focused on simple cases, like a single point, or a hyperplane defined

by a constant balance. In this case, the deviation from equilibrium is naturally identified with the difference between the balance and the equilibrium constant. Hardy-Weinberg equilibrium illustrates this case. Stoichiometry of olivines is a case in which the equilibrium locus is a wrapped hyper-surface. A linearisation is proposed so that measuring deviation from equilibrium is reduced to the linear case. Deviation from equilibrium is measured by the departure of a log-ratio from the equilibrium constant.

Acknowledgements JJE and VPG were supported financially by Ministerio de Economía y Competitividad (MINECO/FEDER (Spain), MTM2015-65016-C2-1-R) and AB by University of Florence (Italy) funds 2020.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London. 416 p. (1986).
2. Aitchison, J.: On criteria for measures of compositional difference. Math Geol **24**(4), 365–379 (1992).
3. Aitchison, J.: A concise guide to compositional data analysis. Unpublished techreport, http://ima.udg.edu/Activitats/CoDaWork05/ (2005).
4. Egozcue, J. J. and V. Pawlowsky-Glahn: Compositional data: the sample space and its structure. TEST **28**(3), 599–638 (2019).
5. Egozcue, J. J., V. Pawlowsky-Glahn, and A. Buccianti: Distances to compositional equilibrium. J Geochem Explor **227**, 106793, (2021).
6. Gard, M., D. Hasterock, and J. A. Halpin: Global whole-rock geochemical database compilation. Earth Syst. Sci. Data **11**, 1553–1566 (2019).
7. Gini, C.: Measurement of inequality of incomes. Econ J **31**(121), 124–126 (1921).
8. Hardy, G. H.: Mendelian proportions in a mixed population. Science **28**, 49–50 (1908).
9. Martín-Fernández et al. Advances in principal balances for compositional data. Math Geosci **50**, 273–298 (2018).
10. Pawlowsky-Glahn, V. and J. J. Egozcue: Geometric approach to statistical analysis on the simplex. SERRA **15**(5), 384–398 (2001).
11. Pawlowsky-Glahn et al.: Modeling and analysis of compositional data. Statistics in practice. Wiley, UK. 272 pp. (2015).
12. Weinberg, W., Über den Nachweis der Vererbung beim Menschen. Jahreshefte d. Vereins vaterl. Naturk. in Württemb. **64**, 369–382 (1908).

The knockoff filter for FDR control in robust ZeroSum regression in microbiome analysis

Il filtro knockoff per il controllo del tasso di false scoperte nella regressione robusta ZeroSum per l'analisi del microbiota

Gianna Serafina Monti and Peter Filzmoser

Abstract In this contribution we consider the knockoff filter in robust ZeroSum regression, an approach to high-dimensional regression with compositional covariates based on a class of shrinkage estimators for least trimmed squares regression in combination with elastic-net penalty [8]. The goal is to select a set of relevant features which have a nonzero effect on the response, and, at the same time, controlling for the false discovery rate. The proposed methodology is particularly useful in the analysis of microbiome compositional data.

Abstract *In questo contributo consideriamo il filtro knockoff nella regressione robusta ZeroSum, un approccio alla regressione ad alta dimensionalità con covariate composizionali [8]. L'obiettivo è selezionare un insieme di variabili rilevanti che abbiano un effetto non nullo sulla risposta e, allo stesso tempo, tenere sotto controllo il tasso di false scoperte. La metodologia proposta è particolarmente utile nell'analisi dei dati composizionali del microbioma.*

Key words: FDR control, knockoff filter, log-contrast model, microbiome.

1 Introduction

The analysis of human microbiome has garnered great attention in recent research thanks to next-generation sequencing technologies which allow to collect a huge amount of data with relatively low costs.

Gianna Serafina Monti
Department of Economics, Management and Statistics, University of Milano Bicocca, e-mail: gianna.monti@unimib.it

Peter Filzmoser
Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology
e-mail: P.Filzmoser@tuwien.ac.at

The sequencing reads data, usually collected into operational taxonomic units (OTUs), are commonly normalized due to the high variations of the microbial abundances among samples, inducing relative abundances. The relevant information for each sample is contained into the ratios between components, thus the microbiome data has essentially the characteristics of compositional data [1].

One of the main interest in microbiome analysis is to identify microbial taxa that are meaningfully associated with an outcome of interest. The high-dimensional settings and the compositional nature of microbiome data stimulated several proposals of variable selection. Recently [8] developed a robust and sparse method for variable selection by means of the least trimmed squares regression approach.

2 Zerosum regression

Let $\mathbf{y} \in \mathbb{R}^n$ denote the response vector, $\mathbf{X}$ the matrix of compositional covariates, $\mathbf{X} = [x_{ij}]_{1\le i\le n; 1\le j\le p}$, whose rows are normalized to a unit sum, with the implicit assumption that $x_{ij} > 0$, and $\mathbf{Z} = [z_{ij} = \log(x_{ij})]_{1\le i\le n; 1\le j\le p} \in \mathbb{R}^{n\times p}$ the log-transformed matrix of $\mathbf{X}$. The symmetric log-contrast regression model [2, 7] is defined as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ st } \sum_{j=1}^{p} \beta_j = 0, \tag{1}$$

where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$ is the error component. Model (1) is also known as ZeroSum regression, due to the constraint of the $\boldsymbol{\beta}$ vector. In an high-dimensional setting, namely $p \gg n$, the variable selection of a log-contrast model is conducted via penalized regression [7].

$$\hat{\boldsymbol{\beta}}_{\text{ZS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left(\frac{1}{n} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right), \text{ st} \sum_{j=1}^{p} \beta_j = 0, \tag{2}$$

where $\lambda > 0$, is the regularization parameter, which calibrates the sparseness, and $\|\cdot\|_2$ and $\|\cdot\|_1$ indicate the ℓ_2 and ℓ_1 norm, respectively.

To cope with the presence of vertical and horizontal outilers, [8] proposed the RobZS estimator, a robust version of the penalized ZeroSum regression model (2), defined as

$$\hat{\boldsymbol{\beta}}_{\text{RobZS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \underset{\substack{H\subseteq\{1,\dots,n\}:\\ |H|=h}}{\arg\min} \left(\sum_{i\in H} (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + h\lambda P_\alpha(\boldsymbol{\beta}) \right), \text{st} \sum_{j=1}^{p} \beta_j = 0\,, \tag{3}$$

where $P_\alpha(\boldsymbol{\beta}) = \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right)$ is the elastic-net penalty, $\alpha \in [0, 1]$ is a tuning parameter which balances the ℓ_2 and ℓ_1 penalty. H is an outlier-free subset of the set

of all indexes $\{1,2,\ldots,n\}$, and $|H|$ denotes the cardinality of the set H. An analog of the fast LTS algorithm [3, 6] was proposed to fit the model (3).

The RobZS estimator fulfills desirable compositional properties such as scale invariance, permutation invariance, and selection invariance.

3 Compositional knockoff filter

The penalized regression models described in Section 2 select a set of relevant taxa, but there they have no guarantee on the false discoveries.

[9] suggested to consider the knockoff filter [4, 5] to control the fase discovery rate (FDR), i.e. the expected fraction of false discoveries among all discoveries, and proposed the compositional knockoff filter (CKF) to address the compositional nature of the data involved. CKF is a two-steps procedure: the compositional screening procedure to select the relevant features under the zerosum constraint, and the controlled step bringing off via the knockoff filter methodology.

In our contribution we propose a robust version of the CKF (RobCKF). More precisely we suggest to implement a robust screening procedure in the fist step, achieved recurring to the least trimmed squared framework by means of the estimator (3), and to a use robust versions of the elastic net estimator for linear regression [6] in the knockoff regression step, as the augmented design matrix is no longer compositional.

Numerical simulation studies demonstrate the ability of RobCKF in controlling the nominal FDR with respect to other existing methods in presence of contaminated data.

References

1. Aitchison, J., The Statistical Analysis of Compositional Data. Caldwell, NJ: Blackburn Press (2003).
2. Aitchison, J., Bacon-Shone, J., Log contrast models for experiments with mixtures. Biometrika **71**(2), 323-330 (1984).
3. Alfons, A., Croux, C., Gelper, S., Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann Appl Stat, **7**(1), 226-248 (2013).
4. Barber, R. F., Candès, E. J., Controlling the false discovery rate via knockoffs. Ann. Stat. **43**(5), 2055-2085 (2015).
5. Candès, E., Fan, Y., Janson, L., Lv, J., Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J R Stat Soc Series B Stat Methodol **80**(3), 551-577 (2018).
6. Kurnaz, F. S., Hoffmann, I., and Filzmoser, P., Robust and sparse estimation methods for high-dimensional linear and logistic regression. Chemometr Intell Lab **172**, 211-222 (2018).

7. Lin, W., Shi, P., Feng, R., and Li, H., Variable selection in regression with compositional covariates. Biometrika **101**(4), 785-797 (2014).
8. Monti, G. S. and Filzmoser, P., Sparse least trimmed squares regression with compositional covariates for high-dimensional data. Bioinformatics **37**(21), 3805-3814 (2021).
9. Srinivasan, A., Xue, L., and Zhan, X., Compositional knockoff filter for high-dimensional regression analysis of microbiome data. Biometrics **77**(3), 984-995 (2021).

Solicited Session SS2 - *Monitoring progress towards SDGs: statistical approaches and methods for measuring poverty, inequalities and food insecurity*
Organizer and Chair: Luca Secondi

Poverty indicators adjusted using local price indexes

Indicatori di povertà corretti con indici locali dei prezzi

Marchetti S., Giusti C., Pratesi M. and Biggeri L.

Abstract This paper focuses on the estimation of poverty incidence at sub-regional level in Italy taking into account the different price levels within the country. The local price level is accounted by computing a spatial price index using retail scanner data on sub-regional retail volumes (units) and price for food. Indexes are obtained by a Country Product Dummy model, where we relax the like to like approach as in the International Comparison Program of the World Bank. We compare prices of products belonging to the same type, instead of comparing exactly the same products. Then, price indexes are used to adjust the national poverty line, which is used to estimate poverty incidence at sub-regional level by means of small area estimation methods taht are needed to get reliable estimates.

Abstract *In questo lavoro si presenta metodologia e applicazione della stima di povertà relativa a livello provinciale in Italia, aggiustando la linea di povertà con un'indice spaziale dei prezzi ricavato dai dati "scanner" della grande distribuzione organizzata. L'indice spaziale dei prezzi è ottenuto utilizzando il modello "country product dummy", dove, al contrario di quanto fa la Banca Mondiale, si confrontano i prezzi dei prodotti appartenenti ad una stessa categoria (COICOP 8 digit) - e non si confrontano esattamente gli stessi prodotti. Inoltre, a casua della ridotta dimensione campionaria, per ottenere le stime di povertà a livello provinciale si utilizzano modelli di stima per piccole aree al fine di ottenre stime attendibili.*

Key words: Country-product-dummy model, small area estimation, scanner data, big data

Marchetti Stefano, Giusti Caterina, Pratesi Monica
Dep. of Economics and Management, University of Pisa, Via C. Ridolfi 10, 56124 Pisa (PI) Italy, Dagum ASESD Centre, e-mail: stefano.marchetti@unipi.it, e-mail: caterina.giusti@unipi.it, e-mail: monica.pratesi@unipi.it

Biggeri Luigi
Dep. of Statistics, Informatics, Applications, University of Florence, Viale Morgagni 59, 50134 Firenze (FI) Italy,
Dagum ASESD Centre, e-mail: luigi.biggeri@unifi.it

Marchetti S., Giusti C., Pratesi M. and Biggeri L.

1 Sub-regional spatial price indexes

The aim of this work is to improve the measure of the incidence of relative poverty - a monetary based measure - by taking into account the different price level within the country. First, we use a methodology to compute Spatial Price Indexes (SPIs) at sub-regional level in Italy using retail scanner data on sub-regional retail volumes (units) and price for food and non alcoholic beverages (food for short). The data were provided by Istat/Nielsen within the framework of the H2020 project Makswell (www.makswell.eu). Secondly, we adjust sub-regional poverty incidence using the sub-regional SPIs.

Specifically, we compute SPIs for 103 (out of 107) Italian provinces, by using the scanner data referring to the year 2018 and only to the products (barcodes or Global Trade Item Numbers - GTINs) in food and beverages categories, excluding fresh food. Usually the information on products' quantities is reported in terms of grams and milliliter, but sometimes in units; given that we needed to use comparable prices, we discarded about $17,000$ quotations expressed in units (out of about $630,000$ quotations). We used a Country Product Dummy (CPD) (Laureti and Rao (2018)) model to obtain SPIs, where we model the average price of unit (gr or ml) of 102 groups of products classified by COICOP with 8 digits, for each provinces. The hypothesis is that products (items) in the same COICOP-8-digit group give the same utility. By this way it is easier to apply the CPD model and we avoid the impact on prices that can be observed for products that are popular in some areas (demand is high) and that they are not in other areas (demand is low). Moreover, average prices obtained from scanner are based on real purchases carried out by households, which, according to the HBS, buy at least 50% of food in supermarket.

Let $\bar{p}_{ij}$ be the mean price for COICOP-8-digit j and province i and let r_{ijk} and q_{ijk} be the annual turnover and the total quantity sold respectively of item k belonging to COICOP-8-digit j in province i. These quantities are estimated by Istat using the scanner data and the sampling weights computed according to the survey design (refer to Deliverable 3.2 of the MAKSWELL project for further details). Let u_{ijk} be the quantity of the item ijk in terms of gr or ml. For each item we define its annual price per gr or ml and its relative weights in term of turnover as

$$p_{ijk} = \frac{\frac{r_{ijk}}{q_{ijk}}}{u_{ijk}} \quad w_{ijk} = \frac{r_{ijk}}{\sum_{k=1}^{n_{ij}} r_{ijk}},$$

where n_j is the number of items in the jth COICOP-8-digit aggregation and the ith province. The weighted mean price per gr. or ml. for products in ECOICOP-8-digit j and province i is:

$$\bar{p}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} p_{ijk} w_{ijk}.$$

The CPD model we propose is as follows:

$$\log \bar{p}_{ij} = \alpha_0 + \alpha_i D_i + \beta_j I_j + \varepsilon_{ij}, \quad i = 1, \ldots, 103 \quad j = 1, \ldots, 102, \tag{1}$$

where D_i is a vector equal 1 if the mean price is in province i and 0 otherwise, I_j is equal 1 if the mean price belongs to jth COICOP-8-digits and 0 otherwise, so that β_js account for difference in quality. We assume the error $\varepsilon_{ij} \sim N(0, \sigma^2)$. Moreover, we consider the different level of the turnover between the COICOP-8-digit aggregates by estimating parameters in (1) using weighted least squares, where the weights are computed as the ratio between the total turnover of one aggregate in one province and the total turnover in the province (n_i is the number of items in the ith province)

$$wls_{ij} = \frac{\sum_{k=1}^{n_{ij}} r_{ijk}}{\sum_{k=1}^{n_i} r_{ijk}}.$$

As it is, model (1) is not identified, because the D_is vectors are a linear combination of the constant. Therefore, we impose the constraint $\alpha_1 = 0$ so that α_i is the fixed effect of province i respect to province 1. Once the parameters are estimated, we apply the method in Suits (1984) to use as a reference Italy instead of area 1. In this way, $\hat{\alpha}_i$ represents the estimated fixed effect of province i compared to Italy and $\exp(\hat{\alpha}_i)$ is, finally, the SPI of province i.

The proposed method can be easily extended to produce SPIs related to the first quintile (or any other quantile) of the distribution of the price of each specific product, assuming that poor purchase the cheaper items of the product. To obtain these SPIs we can modify model (1) using the first quintile (weighted) of the p_{ijk}s as target variable instead of $\hat{p}_{ij}$. Figure 1 shows two choropleth maps of SPIs based on the mean (left) and SPIs based on first quintile (quantile 0.2), denoted as SPI($Q_{0.2}$)'s (right).

The results we obtained are somehow expected. Indeed, provinces in the south of Italy show SPIs smaller than 1, while provinces in the north show values greater than 1. However, there are exceptions, provinces in the north-east Alps mountains show SPIs below 1, even if they are close, both considering the mean and the quantile 0.2 of unit prices.

2 Poverty incidence adjusted using sub-regional spatial price indexes

In Italy the official measure of relative poverty incidence is defined as the proportion of persons whose consumption expenditure is below a fixed threshold, called poverty line. The poverty line for an household of two components is defined as the mean per capita consumption expenditure. For households with members different from two, the poverty line is adjusted using the Carbonaro scale (Istat (2010)), which assigns $0.6, 1.33, 1.63, 1.90, 2.16, 2.40$ to households with $1, 3, 4, 5, 6$ and 7 or more components respectively. Poverty incidence estimates in Italy are based on the household budget survey (HBS), which design allows for reliable estimates at regional level. The poverty line is fixed at the national level.

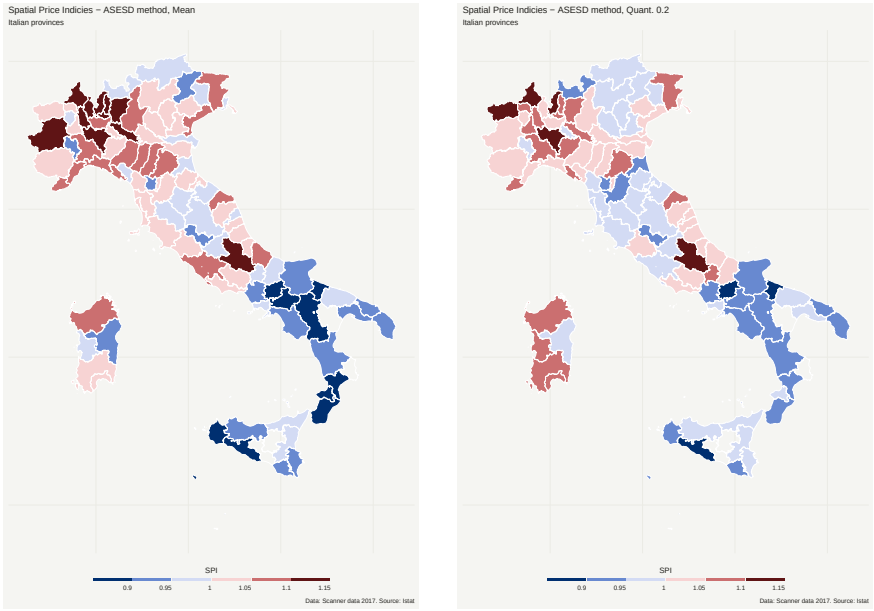


Fig. 1 Choropleth map of SPIs obtained using mean unit prices (left) and quantile 0.2 of unit prices (right).

The use of a national poverty line allows to establish a general scheme of how local areas (e.g. regions or provinces) compare with national standards. However, considering the same poverty line for each area implies an equity concept in which individuals with equal income are assumed to have similar wellbeing regardless of the area where they live. Using the province SPIs we can adjust the national poverty line at the province level, taking into account the different price level within the country.

In this work we adjust the national poverty line using the SPI($Q_{0.2}$) values. Specifically, the national poverty line is adjusted for each province using the SPI($Q_{0.2}$) values opportunely weighted (adapting the idea in Renwick et al. (2014)):

$$nPL_i^* = nPL \times (\lambda_i SPI_i + 1 - \lambda_i) \tag{2}$$

where nPL is the national poverty line, nPL_i^* is the adjusted poverty line for province i, λ_i is the estimated share of food consumption in province i and SPI_i is the SPI($Q_{0.2}$) for province i. The quantities λ_i's are estimated from the HES 2017 as the provincial mean of the ratios between the food expenditure and the total consumption expenditure:

$$\lambda_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} \frac{p_{ij}}{t_{ij}} w_{ij}, \tag{3}$$

where n_i is the sample size in province i, w_{ij} is the survey weight of household j in area i, p_{ij} is the food expenditure of household j in area i and t_{ij} is the total consumption expenditure of household j in area i. The survey weights have been calibrated to sum to the total households at provincial level. Although the λ_i's are estimated at the provincial level – thus possibly unreliable because of small sample size – we judge the direct estimates suitable for our purpose.

Having computed the adjusted nPLs, we then calculated the corresponding direct estimates of the poverty rates. As the variability of the direct estimates was too high (approximately half of the provinces have a CV greater than 30%) we estimated a Fay-Herriot (FH) model with the following auxiliary variables: the ratio between number of taxed persons over the population, and the ratios between the number of persons with *i*. income coming from salary, *ii*. income coming from pensions and *iii*. income lower than 10,000 euros per year, over the number of taxed persons. These data come from the Italian tax agency database 2017. The EBLUPs (Empirical Best Linear Unbiased Predictors) obtained with the FH model showed a gain in efficiency with respect to direct estimates. We obtained a CV smaller than 16% in 37 provinces, while half of the provinces had a CV smaller than 20%. We also computed the EBLUPs without any adjustment of the national poverty line, using the same small area model as for adjusted EBLUPs. Figure 2 reports the comparison of the two set of EBLUPs estimates: as we can see, using the SPI($Q_{0.2}$) to adjust the poverty lines, the HCRs in northern and central provinces slightly decrease.

The results obtained here suggest that the methodology can be extended to include other Spatial Price Indexes, therefore adjusting the national poverty line with other components of households' consumption expenditure. Indeed, our results suggest the products included in the scanner data represent a relevant but still limited share of the total household consumption expenditure, approximately equal to the 20%. Therefore, by including other consumption expenditure components, such as for example the expenditure for the rent, the national poverty line could be adjusted in a more complete manner.

References

1. La differenza nel livello dei prezzi al consumo tra i capoluoghi delle regioni italiane. Istat, Italian national statistical office, Rome, Italy (2010).
2. Laureti, T. and Rao, D.: Measuring spatial price level differences within a country: Current status and future developments. Estudios de economia aplicada 36(**1**), 119-148 (2018).
3. Renwick, T., Aten, B., Figueroa, E., and Martin, T.: Supplemental poverty measure: A comparison of geographic adjustments with regional price parities vs. median rents from the american community survey. Technical report Bureau of Economic Analysis (2014).
4. Suits, D.: Dummy variables: Mechanics v. interpretation. Review of Economics and Statistics 66, 177–180 (1984).

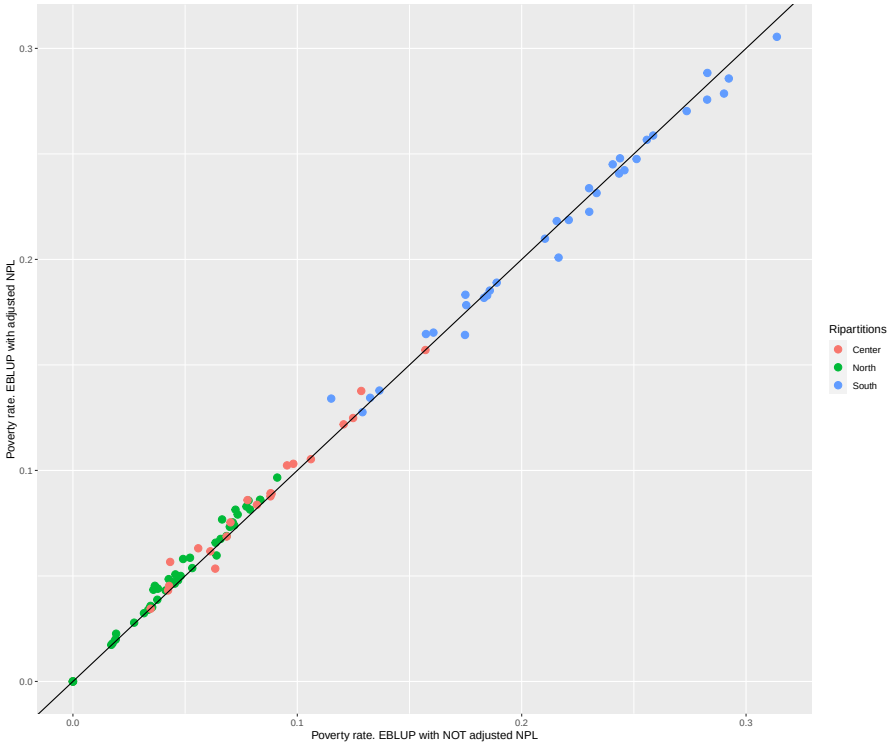


Fig. 2 Poverty rate at provincial level in Italy: provincial EBLUPs estimates using the SPI($Q_{0.2}$) adjusted vs not adjusted national poverty line.

Assessing the targeting of the anti-poverty measure "Reddito di Cittadinanza" using Small Area Estimation methods

Valutazione del targeting del Reddito di Cittadinanza attraverso la stima per piccole aree

Giovanni Tonutti, Gaia Bertarelli, Caterina Giusti and Monica Pratesi

Abstract Sustainable Development Goal 1 calls for the implementation of nationally appropriate social protection systems to contrast poverty. In Italy, a crucial anti-poverty policy is the "Reddito di Cittadinanza" (RdC) introduced in April 2019. In this work we aim at evaluating the targeting of the RdC in 59 local areas represented by the region by degree of urbanisation level in Italy. To measure the local poverty share, we estimate At-Risk-of-Poverty rates and Absolute Poverty rates through the application of Small Area Estimation models. Our results suggest that the RdC shows very heterogeneous targeting performance at the local level, excluding large shares of poor households from the program.

Abstract *L'Obiettivo di Sviluppo Sostenibile 1 richiede l'implementazione di sistemi di protezione sociale adeguati a livello nazionale per contrastare la povertà. In Italia il "Reddito di Cittadinanza" (RdC), introdotto nell'aprile 2019, rappresenta una misura cruciale in tal senso. In questo lavoro valutiamo il targeting del RdC in 59 aree locali rappresentate dai tre gradi di urbanizzazione in ciascuna regione. Per misurare i tassi di povertà, stimiamo il rischio di povertà e la povertà assoluta attraverso l'applicazione di modelli di stima per piccole aree. I nostri risultati suggeriscono che l'RdC mostra un targeting molto eterogeneo a livello locale, escludendo ampie quote di famiglie povere dal programma.*

Key words: Poverty, Targeting Analysis, Small Area Estimation methods

Giovanni Tonutti
Scuola Normale Superiore Pisa, giovanni.tonutti@sns.it

Gaia Bertarelli
Scuola Superiore Sant'Anna, Pisa, gaia.bertarelli@santannapisa.it

Caterina Giusti
Università di Pisa, caterina.giusti@unipi.it

Monica Pratesi
Università di Pisa, monica.pratesi@unipi.it

1 Introduction

In April 2019, the Italian government introduced a national measure of guaranteed minimum income under the name of "Reddito di Cittadinanza" (RdC). RdC represents the largest monetary transfer program to low-income families in the history of the Italian social security system. For the year 2019 alone, total program expenditure was forecasted at €5.6bn, with an estimated cohort of beneficiaries of 1.3m households. Poverty reduction represented the first and key objective of the policy as well as the central theme in the communication campaign leading to the introduction of the measure. Official statistics by the Italian National Statistical Institute (ISTAT) show indeed how the number of households in absolute poverty in Italy had been on the rise over the course of the five years previous to the introduction of the policy. In 2018, the number of families in absolute poverty has reached the figure of 1.8m, with an absolute poverty incidence in the Italian population of 7% (ISTAT 2019), against an estimated total of 1.3m households as potential beneficiaries of the RdC highlighting a gap between the overall cohort of RdC beneficiaries and the total number of families in absolute poverty in Italy (compare INPS 2019). Based on these considerations, this research assesses the extent to which the policy succeeds in targeting support to families in poverty at the local level and which factors related to the local demographic and economic characteristics drive variations in targeting coverage and take-up rates. In addressing these questions, the research will provide the first assessment of the targeting of RdC based on administrative data on its beneficiaries. To capture the geographical heterogeneity in the effects of anti-poverty interventions, we consider as the unit of analysis the degree of urbanization as measured by the DEGURBA classification across the 20 Italian regions. By and large, official poverty indicators are estimated on the basis of surveys collected by national statistical agencies at the national level, and often, due to their limited sample sizes, cannot provide accurate estimates at lower sub-regional units of analysis (Tzavidis et al. 2018). Small Area Estimation (SAE) methods offer the tools to overcome this gap, introducing statistical models that combine the direct estimates obtained from the surveys with error-free administrative covariates to improve the precision of the estimates. While the application of SAE models to study of the geographical distribution of poverty is an established methodology, this work proposes a new application of SAE methods to assess an anti-poverty program. In this paper SAE methods are instrumental to provide the baseline poverty estimates for each of the 59 areas of analysis to successively estimate the targeting performance of the RdC across such areas. The results and conclusions drawn by this research are important for policy makers as they can help them in the design of livelihood policies in the territory where people live. In addition, this paper propose a novel application of SAE methods for assessing local targeting of anti-poverty policies.

2 Data

The analyses presented in this work are based on four main data sources. Estimates for absolute poverty (AP) are based on the HBS data for the year 2017. The survey

provides information on households consumption behaviour. The data-set provides a flag for households living in AP and comprises $\approx$ 17,000 observations. The estimates produced are reliable at regional level but not at the sub-regional due to the limited sample size. Estimates for the at-risk-of-poverty rate (AROP) are based instead on EU-SILC survey collected in 2017. EU-SILC aims at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. The 2017 wave of survey contains information on self-reported income for the year 2016 with a total of 22,200 observations. Finally, information on the number of RdC beneficiaries and the monetary amount of benefit received by municipality was provided by INPS, the Italian Social Security Agency. The data-set identifies the total number of households and individuals in receipt of the scheme as of December 2019.

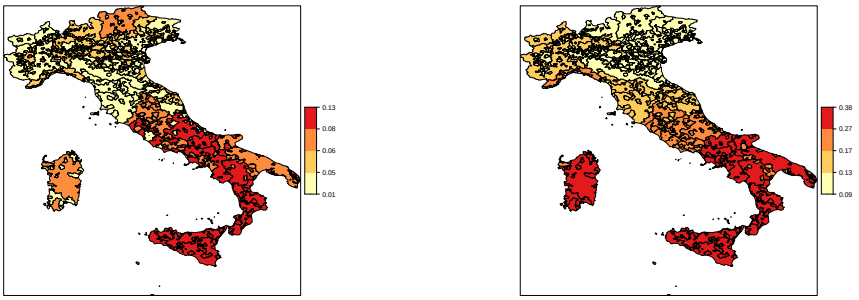
3 Methods

For both AP and AROP estimates, our target indicators are the small area means. The application of SAE models aims at increasing the precision of direct survey estimates through the use of the administrative covariates at DEGURBA $\times$ region level. To this purpose, we apply the Bivariate Fay-Herriot (FH) model (Benevanent and Morales 2016). The FH model and its multivariate transformations are area level models that links direct estimates to area level covariates. In the study of anti-poverty programs, the concept of targeting refers to the attempt by public officials to identify who is poor and then to restrict transfers to those individuals (Hanna and Olken 2018). Data on RdC beneficiaries was made available for this research at the municipality level. This level of aggregation does not allow to identify those recipients of RdC who can be considered as not poor. As such, the most meaningful targeting indicator to be applied in this analysis is the Coverage Rate (CR) metric (Coady et al. 2004) Defining by D_{ij} an indicator variable that takes value 1 if unit j living in area i was beneficiary of the RdC and by c_{ij} and y_{ij} the unit consumption and income measure, and with t_{ij} and t the corresponding poverty lines, two CRs can be defined as following:

$$CR_{iAP} = \frac{\sum_{j=1}^{N_i} D_{ij} \cdot \mathrm{I}(c_{ij} \leq t_{ij})}{\sum_{j=1}^{N_i} \mathrm{I}(c_{ij} \leq t_{ij})} \qquad \text{and} \qquad CR_{iAROP} = \frac{\sum_{j=1}^{N_i} D_{ij} \cdot \mathrm{I}(y_{ij} \leq t)}{\sum_{j=1}^{N_i} \mathrm{I}(y_{ij} \leq t)}.$$

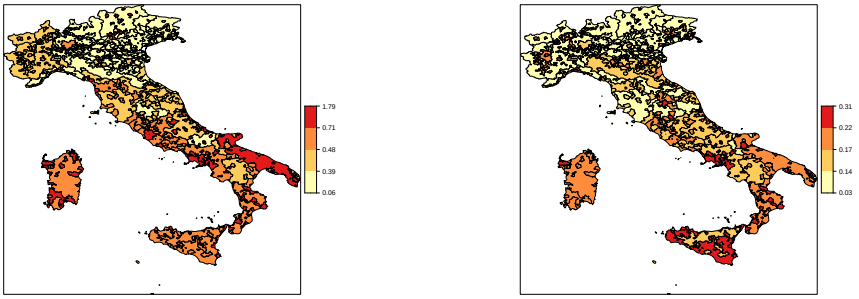
The two measures above correspond to the ratio between the total amounts of households in absolute poverty and at risk of poverty who received the RdC living in area i, over the corresponding total amount of households in absolute and relative poverty in area i.

4 Results and Discussion



(a) (b)

Fig. 1: Estimates of absolute poverty (panel a) and of the AROP (panel b) for the 59 degrees of urbanisation across 20 regions in Italy.



(a) (b)

Fig. 2: Coverage rate of RdC estimated on AP (CR_{AP} - panel a) and on AROP (CR_{AROP} - panel b) for the 59 degrees of urbanisation across Italy 20 regions.

Small area estimations are employed to improve the precision of direct estimates from both HBS and EU-SILC, surveys designed to provide reliable information at higher geographical levels. To assess gains in the accuracy of our estimates we compare the coefficient of variations of bivariate FH model with those of the respective direct estimates. In this analysis, the application of SAE methods brings considerable gains to the precision of estimates as illustrated in Table 1. The bivariate FH model reduces the number of areas with CV estimates above the 33.3% threshold by more than three times, compared to the direct estimates of absolute poverty, leaving

only 7 areas with an uncertainty of estimation too high to be considered as reliable. By contrast, the bivariate FH estimates of AROP show CVs all below the 16.5% threshold. The difference in precision between the two estimates stems from three main reasons.

		<16.5%	16.5-33.3%	>33.3%
AP	Direct	5	32	22
	FH bivariate	11	41	7
AROP	Direct	33	22	2
	FH bivariate	59	0	0

Table 1: Comparison of the coefficients of variation of absolute poverty and AROP estimates.

Figure 1 shows the distribution of poverty for the 59 degree of urbanisation across Italy 20 regions for both AP and AROP indicators. Both maps show a clear distinction in poverty incidence across the country's three main areas of north, centre and south: higher poverty incidence characterise southern areas. The AP index ranges from a maximum of 13.38% for the rural areas of Molise (South), to 1.11% in the sub-urban areas of Trentino-Alto Adige (North). The AROP ranges from 37.62% in sub-urban Sicilian areas (South) to 9.13% in sub-urban Friuli-Venezia Giulia (North). These findings reflect the country long-lasting economic dualism. While following a clear north-south divide, the geographical distribution of absolute poverty incidence shows variation within the three main geographical areas. The second main consideration is related to the within region heterogeneity in the incidence of AP, in contrast with a rather homogeneous within-region distribution of the AROP indicator. When poverty is measured on consumption, there seems to be greater variations within the same region across different degrees of urbanisation. The considerations highlighted so far are the result of differences in the definition of the poverty indicators considered by the analysis. As discussed in Section 2, AP is estimated on the basis of consumption behaviour based on different poverty lines, varying across Italy three macro-areas and across the size and type of municipality of the survey respondent. Unlike single national poverty threshold, such as the AROP indicator present in the EU-SILC data, this approach allows to capture differences in costs of living. Given the limitations of the data on RdC beneficiaries, which are available at aggregate municipalities level, and the difficulty in excluding non-poor recipients from the overall share, the targeting indicators are likely an overestimate of the true parameter. Figure 2 plots the two CR indicators of the RdC for each of the 59 DEGURBA areas across the 20 Italian regions. We observe a rather heterogeneous distribution of the CR indicators across Italy. The main difference in the comparison of the two indicators is in the width of the range of values. The CR_{AP} indicator ranges from 5.6% in the rural areas of Trentino-Alto Adige to 179.31% in the sub-urban areas of Sardinia. Values of the CR_{AROP} indicator, on the contrary, show a significant narrower range, from 3.31% of rural areas in Trentino-

Alto Adige to 31.07% of urban areas in Sicily. The CR_{AROP} indicator highlights how the vast majority of households identified as at risk of poverty are excluded from the support provided by the RdC. The CR_{AP} indicator, on the contrary, describes a policy with large geographical heterogeneity in its targeting performance, excluding large number of absolute poor households in areas with higher costs of living, and including non-poor households in more affordable ones. Overall, the policy seems to consistently show lower targeting performance in the northern areas of the country, especially in the North East. Approximately all among the bottom 10 areas for both CR_{AP} and CR_{AROP} indicators are in the North. Moreover, if we consider both CR_{AP} and CR_{AROP} metrics, rural areas across Italy present lower targeting performance, irrespective of the three macro-areas considered.

5 Conclusion

In this work we presented a first study on the targeting at the local level of the RdC anti-poverty policy in Italy. The study was based on four main data sources and made used of appropriate SAE techniques to obtain reliable poverty estimates for the 59 local areas of interest. It is essential to implement local level targeting of anti-poverty policies to meet the needs and problems of the territory where people live. The results of this study show an heterogeneous targeting performance of the RdC policy, with a general lower targeting affecting norther regions and rural areas.

References

1. Benavent, R., and Morales, D.: Multivariate Fay–Herriot models for small area estimation. Computational Statistics & Data Analysis **94**, 372-390 (2016). **169**, 321–354 (1995)
2. Coady, D., Grosh, M. E. and Hoddinott, J.: Targeting of transfers in developing countries: Review of lessons and experience (2004)
3. Hanna, R., and Olken, B. A.: Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. Journal of Economic Perspectives **169**, 201-26 (2018).
4. INPS: Reddito/pensione di cittdaninanza e reddito di inclusione. osservatorio statistico, Technical report, Istituto Nazionale Previdenza Sociale (2019)
5. ISTAT: Le statistiche dell'istat sulla poverta'. Anno 2018 (2019)
6. Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T. and Rojas-Perilla, N.: From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society) **181.4**, 927-979 (2018)

Household Consumption and Food Insecurity in Mexico: Covid19 and Sustainable Development

Consumi delle famiglie e insicurezza alimentare in Messico: Covid19 e sviluppo sostenibile

Adrian Vargas-Lopez and Luca Secondi

Abstract Attaining a lower level of food insecurity is crucial for developing countries as its consequences spread wide and deep into specific communities. Covid-19 has magnified the adverse effects of several problems worldwide, including food security. This study investigates the Mexican Households' four food security thresholds using the 2018 and 2020 waves of the National Household Income and Expenditure Survey (ENIGH), which contains the Latin American and Caribbean Food Security Scale (ELCSA). In this research, we assess the differences in the four food security categories with reference to both individual and household variables as well as contextual factors.

Abstract *Raggiungere un livello inferiore di insicurezza alimentare è fondamentale per i paesi in via di sviluppo poiché le conseguenze si estendono in modo ampio e profondo in comunità specifiche. Il Covid-19 ha amplificato gli effetti negativi di diversi problemi in tutto il mondo, inclusa la sicurezza alimentare. Questo studio indaga le quattro soglie di sicurezza alimentare delle famiglie messicane utilizzando i microdati delle indagini 2018 e 2020 del National Household Income and Expenditure Survey (ENIGH), che contiene la scala di sicurezza alimentare dell'America Latina e dei Caraibi (ELCSA). In questa ricerca, valutiamo le differenze nelle quattro categorie di sicurezza alimentare sia con riferimento a variabili individuali e familiari che contestuali.*

Keywords: Food Insecurity, Covid19, Mexico, ELCSA

Adrian Vargas-Lopez, Tecnologico de Monterrey, School of Government and Public Transformation, Mexico City, Mexico; email: a.vargaslopez@tec.mx

Luca Secondi, University of Tuscia, Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF), Via S. Camillo De Lellis, snc, Viterbo, 01100, Italy; email: secondi@unitus.it

1 Introduction

Food security is defined as "*having at all times, physical, social and economic access to sufficient, safe and nutritious food that meets dietary needs and food preferences for an active and healthy life*" (World Food Summit, 1996). On the other end, household food insecurity is a significant threat that targets vulnerable groups (Vilar-Compte et al., 2014). According to figures from FAO, almost 811 million people faced hunger last year.

The physical consequences of all forms of malnutrition intensify problems related to chronic illnesses, obesity and additional forms of maladies (Santana-Cárdenas and López-Uriarte, 2021). People living in food insecurity conditions significantly reduce their quality of life and cut their life expectancy (Hampton, 2007). Thus, reducing the number of people that suffer from food insecurity is morally urgent.

We know global crises intensify problems that individuals face daily (Vilar-Compte et al., 2014; Vilar-Compte et al., 2019). The Covid-19 pandemic is not the exception since several studies suggest that food supply chains were disrupted (Singh et al., 2021). Living in a family heavily hit by the pandemic made things more difficult for each member, where infants suffered the most (Magaña-Lemus et al., 2016).

In this study, we explore the likelihood of being into the four food security thresholds by referring to two different waves of the Mexican Household Income and Expenditure Survey (ENIGH) carried out in 2018 (before) and 2020 (during the Covid-19 pandemic).

The remainder of this paper is organized as follows. In the next section, we briefly describe the food security status in Mexico. Then, in Section 3, we describe the data and briefly mention the type of model we selected. In Section 4, we describe our results, while in Section 5, we draw the main conclusions and further necessary progress of the research.

2 Food insecurity in Mexico

In Mexico, individuals experiencing severe food insecurity are geographically located in some of the poorest regions (Mundo-Rosas et al., 2018). These areas are predominantly rural sites in the southern part of the country. Additionally, of these families, when asked if they speak an indigenous language, most state they do (Mundo-Rosas et al., 2018).

Concerning Mexicans' type of diet, Mundo-Rosas et al., 2019 find that having less healthy diets correlates with harsher food insecurity levels. Moreover, they also find that people with severe food insecurity have remained unchanged from 2012 until 2018, at 43%. Magaña-Lemus et al., 2016 paint a clear picture regarding

the characteristics of the head of the household. Dwellings, where the head of the household is a woman with less education, single or widowed, younger, with a disabled relative, experience higher insecurity levels (Magaña-Lemus et al., 2016). Mora-Rivera and van Gameren, 2021 find that homes with access to remittances improve their food security conditions (Mora-Rivera, J. and van Gameren, E., 2021).

3 Data and Methods

The data we used for the analysis considers two waves retrieved from the National Household Income and Expenditure Survey (ENIGH), a nationally representative survey conducted every two years. Most countries in Latin America measure food insecurity using the Latin American and Caribbean Food Security Scale (ELCSA). These are six questions where families signal if during the past three months they had access to a limited variety of food, whether they skipped a meal, if they had eaten less than they thought they should, if they ran out of food, if they felt hungry but did not eat, and if they had not eaten for a whole day. These questions are asked twice if in the household there are children (i.e., individuals younger than 18). The second time, respondents answer for the infants living in the dwelling (Villagómez-Ornelas, 2014).

The severity of food insecurity is constructed by the number of questions that people answer affirmatively. When households without children answer "Yes" to 5-6 questions, they are *Severely Insecure*. If they answer 3-4 questions affirmatively, they are *Moderately Insecure*; 1-2 questions, *Mildly Insecure*; and, 0 questions, *Secure*. Similarly, each threshold is built for households with and without children. Those homes with children that answer 8-12 questions affirmatively are *Severely Insecure*; 4-7, *Moderately Insecure*; 1-3, *Mildly Insecure*; and 0, *Secure*.

Additionally, the data we include in our model is if people live in an urban or rural condition (1 "Rural" or 0 "Urban"), their level of socioeconomic status (1 "Low", 2 "Medium Low", 3 "Medium High" or 4 "High"), the gender of the head of the household (1 "Male" or 0 "Female"), if they receive government's aid (1 "Yes" or 0 "No"), if the household receives remittances (1 "Yes" or 0 "No"), if they have a form of debt (1 "Yes" or 0 "No"), if they have received donations in the past three months (1 "Yes" or 0 "No"), if they live in a household with children (1 "Yes" or 0 "No"), the type of diet consumed at home to meet basic needs (1 "Poor", 2 "Bordering" or 3 "Acceptable"), and the region they belong (i.e., eight regions in total). Where traditionally, regions 6 and 7 are the ones with higher levels of poverty.

To analyze food insecurity in households, we used a multinomial logistic regression. The preliminary analysis considers the variables described in the data section as the vector of X_j independent variables and the four possible thresholds of food insecurity as the *k* categorical outcomes. Furthermore, we stick to the traditional approach of multinomial logistic regressions shown in Greene, 2012. Since the

interpretation of the coefficients from multinomial logit is not straightforward because they are relative to the base outcome, we evaluated the effect of covariates through Marginal Effects (ME) of changing their values on the probability of observing an outcome.

4 Results

Some of the preliminary results from Tables 1 and 2 show the average marginal effects of food insecurity conditions for 2018 and 2020.

Table 1: Average Marginal Effects by Food Insecurity Condition for 2018

	Secure	***Mild***	***Moderate***	***Severe***
Urban (ref.)	-	-	-	-
Rural	-0.007	0.017***	-0.003	-0.007***
	(0.005)	(0.004)	(0.003)	(0.003)
Socioeconomic (Low) (ref.)	-	-	-	-
Socioeconomic (MedLow)	0.076***	-0.030***	-0.023***	-0.024***
	(0.006)	(0.004)	(0.004)	(0.003)
Socioeconomic (MedHigh)	0.170***	-0.066***	-0.050***	-0.054***
	(0.008)	(0.006)	(0.005)	(0.004)
Socioeconomic (High)	0.255***	-0.100***	-0.076***	-0.079***
	(0.009)	(0.007)	(0.005)	(0.004)
Head HH (Female) (ref.)	-	-	-	-
Head Household (Male)	0.047***	-0.006*	-0.016***	-0.025***
	(0.004)	(0.003)	(0.003)	(0.002)
Government beneficiary	-0.095***	0.048***	0.029***	0.019***
	(0.004)	(0.003)	(0.003)	(0.002)
Receives remittances	0.028***	0.000	-0.013***	-0.016***
	(0.007)	(0.006)	(0.004)	(0.004)
Is in debt	-0.063***	0.020***	0.032***	0.011***
	(0.007)	(0.005)	(0.004)	(0.004)
Receives donations	-0.079***	0.036***	0.025***	0.017***
	(0.005)	(0.004)	(0.003)	(0.003)
Household with infants	-0.069***	0.026***	0.049***	-0.007***
	(0.004)	(0.003)	(0.002)	(0.002)
Diet (Poor) (ref.)	-	-	-	-
Diet (Bordering)	-0.016	0.093***	0.010	-0.086***
	(0.026)	(0.017)	(0.020)	(0.024)
Diet (Acceptable)	0.251***	0.064***	-0.063***	-0.252***
	(0.024)	(0.016)	(0.019)	(0.023)
Region 1 (ref.)	-	-	-	-
Region 2	-0.071***	0.029***	0.031***	0.011***
	(0.007)	(0.006)	(0.005)	(0.004)
Region 3	-0.094***	0.056***	0.026***	0.011***
	(0.007)	(0.006)	(0.004)	(0.004)
Region 4	0.005	-0.019***	0.002	0.012***
	(0.007)	(0.005)	(0.004)	(0.004)
Region 5	0.028***	-0.008*	0.009**	0.027***
	(0.006)	(0.005)	(0.004)	(0.003)
Region 6	-0.098***	0.019***	0.038***	0.042***
	(0.007)	(0.006)	(0.005)	(0.004)
Region 7	-0.146***	0.077***	0.042***	0.028***
	(0.008)	(0.007)	(0.005)	(0.004)
Region 8	-0.030***	0.004	0.013***	0.013***
	(0.007)	(0.005)	(0.004)	(0.004)
N	74,647	74,647	74,647	74,647

Household consumption and Food Security in Mexico
Notes: Ref.- Reference category; SE in parenthesis; * p<0.10, ** p<0.05, *** p<0.01

Table 2: Average Marginal Effects by Food Insecurity Condition for 2020

	Secure	***Mild***	***Moderate***	***Severe***
Urban	-	-	-	-
Rural	0.023***	-0.000	-0.009***	-0.014***
	(0.004)	(0.004)	(0.003)	(0.002)
Socioeconomic (Low)	-	-	-	-
Socioeconomic (MedLow)	0.084***	-0.035***	-0.026***	-0.024***
	(0.005)	(0.004)	(0.003)	(0.003)
Socioeconomic (MedHigh)	0.188***	-0.073***	-0.061***	-0.054***
	(0.007)	(0.006)	(0.004)	(0.004)
Socioeconomic (High)	0.274***	-0.112***	-0.082***	-0.081***
	(0.008)	(0.006)	(0.005)	(0.004)
Head Household (Female)	-	-	-	-
Head Household (Male)	0.035***	-0.004	-0.015***	-0.016***
	(0.004)	(0.003)	(0.002)	(0.002)
Government beneficiary	-0.027***	0.024***	0.008***	-0.005***
	(0.004)	(0.003)	(0.002)	(0.002)
Receives remittances	0.011	0.011*	-0.011**	-0.011***
	(0.007)	(0.006)	(0.004)	(0.004)
Is in debt	-0.073***	0.031***	0.024***	0.018***
	(0.006)	(0.005)	(0.004)	(0.003)
Receives donations	-0.081***	0.030***	0.028***	0.024***
	(0.004)	(0.003)	(0.003)	(0.002)
Household with infants	-0.084***	0.030***	0.054***	-0.000
	(0.003)	(0.003)	(0.002)	(0.002)
Diet (Poor)	-	-	-	-
Diet (Bordering)	0.089***	0.056***	-0.030	-0.115***
	(0.025)	(0.020)	(0.022)	(0.025)
Diet (Acceptable)	0.365***	0.036*	-0.110***	-0.291***
	(0.023)	(0.019)	(0.021)	(0.024)
Region 1	-	-	-	-
Region 2	-0.075***	0.042***	0.036***	-0.003
	(0.007)	(0.006)	(0.004)	(0.003)
Region 3	-0.093***	0.050***	0.033***	0.010***
	(0.007)	(0.006)	(0.004)	(0.004)
Region 4	0.047***	-0.036***	-0.010***	-0.001
	(0.006)	(0.005)	(0.004)	(0.004)
Region 5	0.021***	-0.033***	0.001	0.010***
	(0.005)	(0.004)	(0.003)	(0.003)
Region 6	-0.085***	0.020***	0.039***	0.026***
	(0.007)	(0.006)	(0.004)	(0.004)
Region 7	-0.109***	0.061***	0.036***	0.012***
	(0.008)	(0.006)	(0.005)	(0.004)
Region 8	0.026***	-0.021***	-0.001	-0.004
	(0.006)	(0.005)	(0.004)	(0.003)
N	89,006	89,006	89,006	89,006

Notes: Ref.- Reference category; SE in parenthesis; * p<0.10, ** p<0.05, *** p<0.01

The analysis of the two ENIGH waves led us to a picture of household food (in)security in Mexico and changes that occurred with the COVID-19 pandemic. As a general result, we found that food insecurity conditions remained similar before and during the pandemic for socioeconomic status. People living in a high economic power household still maintain higher chances of food security (and vice versa). However, specific findings deserve to be mentioned and further investigated. First, people living in rural zones were less likely to become food insecure during the Covid-19 period than people living in urban areas. Second, receiving government's aid has helped to stave off severe food insecurity. Third, maintaining a close to or

acceptable diet confirmed in both 2018 and 2020 (during the pandemic) is an essential driver of reducing household food insecurity, holding other variables constant. Lastly, the regional effect - proxied at this stage by the dummy variables in the model - shows that the probability of being severely food insecure in 2020 measured through ME was lower than 2018 in regions 5 and 7, while it remained constant for region 1, ceteris paribus.

5 Conclusions

Traditional public policies designed to reverse the effects of food insecurity have to account for the multiple variables operating at different levels. In this study, we begin to explore and understand how the regions have different food insecurity levels and how the context where individuals live can explain food insecurity levels. It is worth noting that these are preliminary results, and further analysis is needed in order to disentangle correctly, through a multilevel approach, the variability of food insecurity at least into individual/households and state/regional levels.

References

1. Greene, W.: Econometric Analysis. 7th Ed. Upper Saddle River, NJ: Prentice Hall (2012)
2. Hampton, T.: Food insecurity harms health, well-being of millions in the United States. JAMA **298**, 1851–1853 (2007)
3. Magaña-Lemus, D., Ishdorj, A., Parr Rosson III, C., Lara-Álvarez, J.: Determinants of household food insecurity in Mexico. Agricultural and Food Economics **4**, 1–20 (2016)
4. Mora-Rivera, J., van Gameren, E.: The impact of remittances on food insecurity: Evidence from Mexico. World Development **140**, 105349 (2021)
5. Mundo-Rosas, V., Vizuet-Vega, N., Martínez-Domínguez, J., Morales-Ruán, M., Pérez-Escamilla, R., Shamah-Levy, T.: Evolución de la inseguridad alimentaria en los hogares mexicanos: 2012-2016. Salud Pública México **60**, 309–318 (2018)
6. Mundo-Rosas, V., Unar-Munguía, M., Hernández-F, M., Pérez-Escamilla, R., Shamah-Levy, T.: La seguridad alimentaria en los hogares en pobreza en México: una mirada desde el acceso, la disponibilidad y el consumo. Salud Pública México **61**, 866–875 (2019)
7. Santana-Cárdenas, S., López-Uriarte, J.: Food insecurity and quality of life in Mexico: a review of studies with a qualitative approach. Journal de Ciencias Sociales **16,** 4–20 (2021)
8. Singh, S., Kumar, R., Panchal, R., Kumar Tiwari, M.: Impact of Covid-19 on logistics systems and disruptions in food supply chain. International Journal of Production Research **59**, 1993–2008 (2021)
9. Vilar-Compte, M., Sandoval-Olascoaga, S., Bernal-Stuart, A., Shimoga, S., Vargas-Bustamante, A.: The Impact of the 2008 financial crisis on food security and food expenditures in Mexico: a disproportionate effect on the vulnerable. Public Health Nutrition **18**, 2934–2942 (2014)
10. Vilar-Compte, M., Gaitán-Rossi, P., Flores, D., Pérez-Cirera, V., Teruel, G.: How do context variables affect food insecurity in Mexico? Implications for policy and governance. Public Health Nutrition **23**, 2445–2452 (2019)
11. Villagómez-Ornelas, P., Hernández-López, P., Carrasco-Enriquez, B., Barrios-Sánchez, K., Pérez-Escamilla, R., Melgar-Quiñónez, H.: Validez estadística de la Escala Mexicana de Seguridad Alimentaria y la Escala Latinoamericana y Caribeña de Seguridad Alimentaria. Salud Pública de México **56**, s5–s11 (2014)
12. World Food Summit, Rome Declaration on World Food Security (1996)

Session of free contributes SCL1 – *Big Data, Proximity data, Multi way data*
Chair: Donatella Vicari

Forecasting Traffic Flows with Complex Seasonality using Mobile Phone Data

Previsione di flussi di traffico a stagionalità complessa con dati di telefonia mobile

Rodolfo Metulini and Maurizio Carpita

Abstract In the era of big data, the ones extracted from mobile phones increase the potentiality for forecasting the amount of traffic flows in a specific area and in a specific time interval. Traffic flows among two regions, however, present a peculiar time series structure, where the daily and the weekly periods are strongly pronounced. For a good prediction performance, one needs to consider a time series modelling structure that takes into account for this kind of complex seasonality. Using one year of mobile phone traffic flows data, retrieved at one hour intervals, in this short paper we aim at forecasting with Harmonic Dynamic Regression models the flows in the strongly urbanized and flooding risk area of the Mandolossa, at the western outskirt of Brescia.
Abstract *Big data estratti dai telefoni cellulari incrementano le potenzialità nella previsione dei flussi di traffico in una determinata area e in un determinato intervallo di tempo. I flussi di traffico presentano una peculiare struttura di serie storica, dove le periodicità giornaliere e settimanali sono fortemente pronunciate. Per una buona "performance" previsiva, è necessario quindi considerare dei modelli che tengano conto di questo tipo di stagionalità complessa. Utilizzando un anno di dati sui flussi di traffico di telefonia mobile, recuperati ad intervalli di un'ora, in questo articolo ci poniamo l'obiettivo di prevedere, con un modello appropriato, i flussi nell'area urbanizzata e a rischio di alluvione della Mandolossa, alla periferia occidentale di Brescia.*

Key words: Origin-destination data, high frequencies time series, flood risk

Rodolfo Metulini
Department of Economics and Statistics (DISES), University of Salerno, Via Giovanni Paolo II, 132 - 84084 Fisciano (SA), Italy, e-mail: rmetulini@unisa.it

Maurizio Carpita
Data Methods and Systems Statistical Laboratory (DMS StatLab), Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, 25122, Brescia, Italy e-mail: maurizio.carpita@unibs.it

1 Introduction

Monitoring and forecasting people mobility is a relevant aspect for metropolitan areas, and in smart cities Information and communication technologies (ICT) with big data are massively used to support the optimization of traffic flows and the study of urban systems [1, 3, 4]. The new technology of mobile phone network data suits with the aim of producing dynamic information on people's presences [6] and movements [7] that can be used also to develop dynamic exposure to flood risk maps for areas with hydrogeological criticity, as proposed in [2]. From a prevention perspective, this could make the identification of preferential traffic flows possible, thus evidencing potential risks during inundation onsets or emergency situations.

Using data provided by Olivetti (*www.olivetti.com/en/iot-big-data*) and FasterNet (*www.fasternet.it*) for the MoSoRe Project 2020-2022 on the flow of mobile phone signals of TIM (*Telecom Italia Mobile*) users among different *census areas* (ACE of ISTAT, the *Italian National Statistical Institute*), recorded on hourly basis for twelve months from September 2020 to August 2021, in this Project we aim at modelling such a flows for the *Mandolossa* area (a critical zone with flood episodes in the north-west of the city of Brescia, Italy) to predict the amount of traffic flows in the context of smart cities emergency management plans.

As traffic flows from TIM mobile phone data show strong daily and weekly patterns, the Harmonic Dynamic Regression (HDR) model with multiple seasonal periods and ARIMA error proposed in [5] suits for our purposes. The short paper is structured as follows: Section 2 presents data, pre-processing and preliminary evidences, Section 3 explain the HDR model adopted and the preliminary results. Section 4 concludes the short paper.

2 Data, pre-processing and preliminar evidences

The original TIM mobile phone data flows are a square origin-destination (OD) matrix of dimension $N \times N$, with $N = 235$ being the number of ACE in the Province of Brescia, retrieved at hourly time intervals from September 2020 to August 2021, so the length of the time series is 24×365. The ACE of interest are four (*Brescia Mandolossa*, *Gussago*, *Cellatica*, *Rodengo Saiano*), which intersect with the identified flooding-risk area, as reported in the map of Fig. 1. We identified other 38 neighboring ACE (aggregated, as represented in Fig. 1, in 4 macro areas), which fulfil the criteria of having a minimum outflow from or to the four ACE of the Mandolossa: the total flows counted between the 4 Mandolossa's ACE and the 38 selected neighboring ACE counts for about the 84% of the total outflows from or to the Mandolossa's ACE.

We consider the time series of the flows from the ACE of *Brescia Mandolossa* to the ACE of *Gussago* (green and yellow respectively on the map

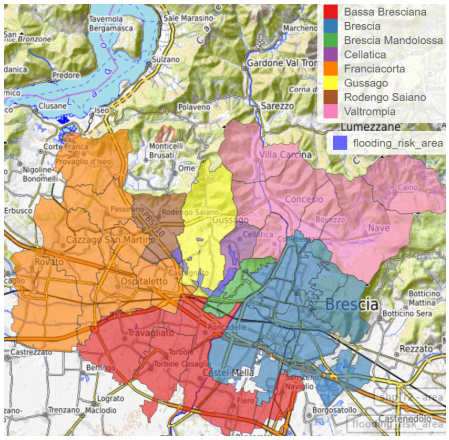


Fig. 1 Map of flooding risk area, ACE of Mandolossa and neighboring macro-areas.

of Fig. 1) and we study the partial autocorrelation function (PACF): Fig. 2 (top right) displays this autocorrelation till a time lag of 168 (one week), and shows that it is very strong for the first two lags (positive for the first and negative for the second); also a strong negative partial autocorrelation emerges among the volume of traffic flows in the same hour interval of a different day of the week. These evidences suggest to model the time series dependence by accounting for daily and weekly seasonality. To explore further the seasonality issue, Fig. 2 (bottom) shows the additive decomposition of the time series of flows from Gussago to Brescia Mandolossa (left) and vice-versa (right), obtained using the Seasonal-Trend decomposition using LOESS (STL) for daily and weekly seasonality [5]. According to the height of the related grey bars on the left of each graph, the daily pattern (seasonal_24) is the most important, whereas the importance of the trend is smaller than the other components. Furthermore, it emerges the presence of some outliers in the residuals (Fig. 2, bottom charts): to avoid negative effects on the estimated models, data have been replaced with $\mu + 3 \cdot \sigma$ (where μ and σ are the mean and the standard deviation of the time series) all flows larger than this cut-off.

3 Modelling the complex seasonal patterns

After data pre-processing, the two time series of flows have been modelled using the Harmonic Dynamic Regression (HDR) with multiple seasonal periods represented by *sin* and *cos* functions of a Fourier basis, and a non seasonal Auto Regressive Integrated Moving Average (ARIMA) error [5]:

$$Flow = \alpha + fourier\_day(k_d) + fourier\_week(k_w) + month + \epsilon\_ARIMA(p, d, q)$$

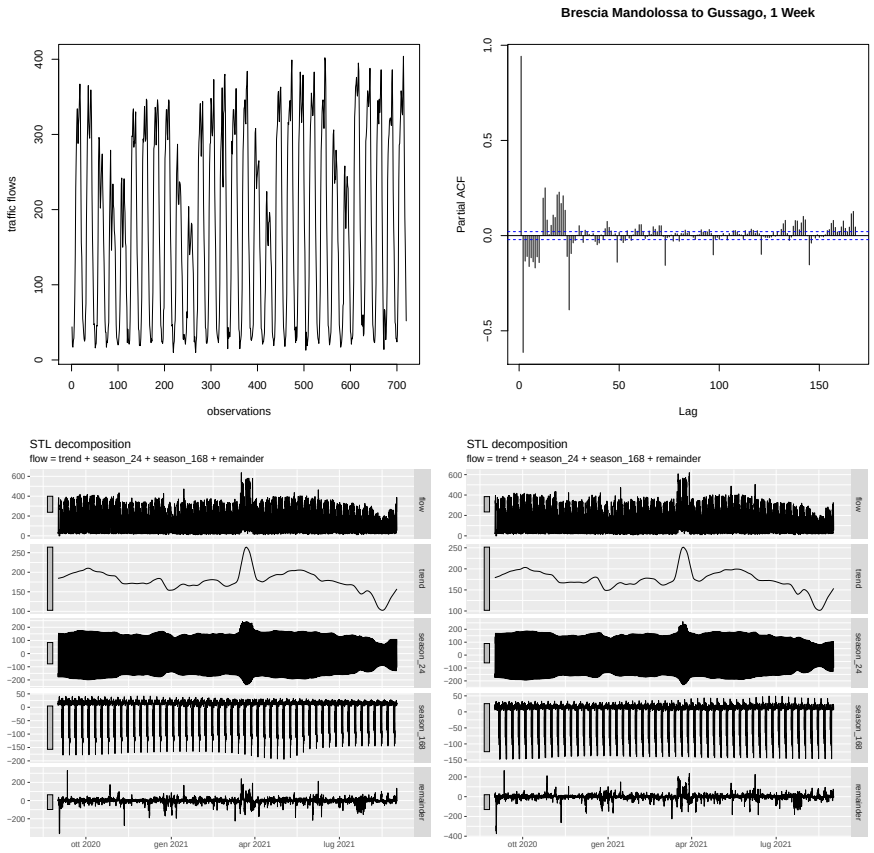


Fig. 2 Traffic flows from Brescia Mandolossa to Gussago in April 2021 (top left) and PACF (top right) with one week of hourly lags. STL decomposition with trend, daily, weekly and error components for one year of traffic flows from Gussago to Brescia Mandolossa (bottom left) and from Brescia Mandolossa to Gussago (bottom right).

where k_d and k_w (the number of Fourier bases for each seasonal component) is selected by minimising the Akaike Information Criterion (AIC). More in details, we choose that numbers by a two step approach: (1) k_d has been selected by minimizing the AIC for the model with no ARIMA error; (2) k_w has been selected minimizing the AIC for the model with k_d chosen in the first step and no ARIMA error. As a result, for both flows' directions we obtain $k_d = 7$ Fourier bases for the daily component and $k_w = 4$ Fourier bases for the weekly component; it worth noting that the number of parameters to be estimated for each Fourier basis is two (*sin* and *cos*). We also include dummies for the months to control for the possible presence of changes in levels (e.g., higher traffic flows in march, as shown in STL charts of Fig. 2).

Results are reported in Tab. 1: daily and weekly Fourier bases as well as most of the monthly dummies show statistical significance and, for both models, error is a stationary autoregressive process of order $p = 5$. Finally, in Fig. 3 an example of one day forecast with confidence bands estimated with the HDR model and the boostrap are presented for both flows.

Table 1 Estimated Harmonic Dynamic Regression Models with ARIMA errors.

Regressors	Brescia Man. → Gussago	Gussago → Brescia Man.
Intercept	160***	160***
$fourier\_day(1)_{cos;sin}$	-132.0*** ; -62.8***	-135.0*** ; -54.7***
$fourier\_day(2)_{cos;sin}$	-9.6*** ; -12.1***	-21.2*** ; -23.5***
$fourier\_day(3)_{cos;sin}$	13.5*** ; 24.0***	19.8*** ; 20.3***
$fourier\_day(4)_{cos;sin}$	0.1 ; -4.1***	-2.1*** ; -0.2
$fourier\_day(5)_{cos;sin}$	-8.1*** ; 0.3	-8.9*** ; 0.7
$fourier\_day(6)_{cos;sin}$	3.0*** ; 1.7***	3.6*** ; 0.2
$fourier\_day(7)_{cos;sin}$	1.7*** ; 5.4***	3.8*** ; 5.3***
$fourier\_week(1)_{cos;sin}$	28.5*** ; -6.9***	29.5*** ; -7.7***
$fourier\_week(2)_{cos;sin}$	-15.9*** ; -11.0***	-15.8*** ; 11.8***
$fourier\_week(3)_{cos;sin}$	7.7*** ; -4.6***	8.0*** ; -5.1***
$fourier\_week(4)_{cos;sin}$	-6.7*** ; -4.7***	-7.2*** ; -4.4***
Month (ref: *January*)		
February	12.3	13.7
March	43.7***	44.2***
April	19.1***	20.9***
May	32.4***	35.9***
June	12.4	15.2
July	-11.4	-9.6
August	-30.3***	-30.0***
September	30.2***	32.7***
October	34.1***	37.5***
November	11.0	12.0
December	0.2	1.5
ARIMA(p,d,q)	5,0,0	5,0,0
AIC	78,885	80,093

Notes: Significance levels for t test: . p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.001.

4 Conclusions

In this short paper, after studying the seasonal structure of the traffic flows recorded between two ACE of the Mandolossa area in the north of Brescia (Italy) using the hourly TIM mobile phone data from September 2020 to August 2021, the HDR model with ARIMA error is estimated.

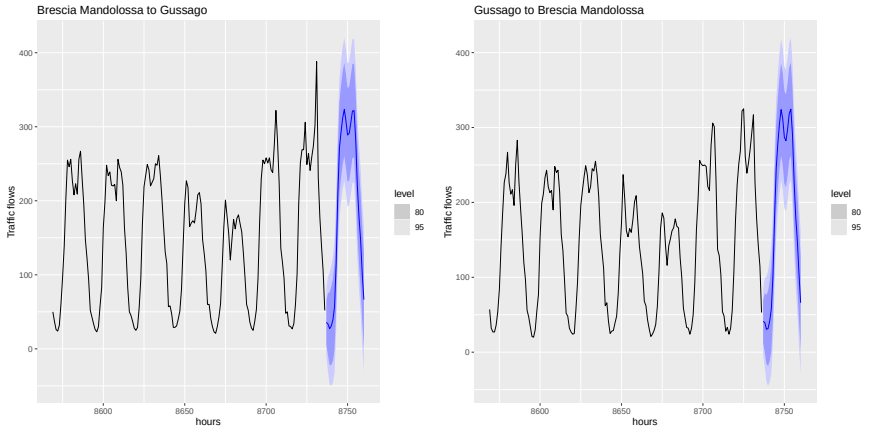


Fig. 3 One day forecast of traffic flows from Brescia Mandolossa to Gussago (left) and from Gussago to Brescia Mandolossa (right) with the HDR models in Tab. 3. Bootstrap confidence bands are reported for the confidence levels of 80% and 95%.

Preliminary results show the statistical significance of daily, weakly and monthly effects, and the opportunity offered by this model to forecast future traffic flows.

Acknowledgements DMS StatLab of the University of Brescia (dms-statlab.unibs.it), cofunded by MoSoRe@UniBS (Infrastrutture e servizi per la Mobilità Sostenibile e Resiliente) Project of Lombardy Region, Italy (CallHub ID 1180965; bit.ly/2Xh2Nfr).

References

1. Albino, V., Berardi, U., D'angelico, R. M.: Smart cities: Definitions, dimensions, performance, and initiatives. Journal of Urban Technology **22(1)**, 3–21 (2015).
2. Balistrocchi, M., Metulini, R, Carpita, M., Ranzi, R.: Dynamic maps of human exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences **20(12)**, 3485–3500 (2020).
3. Benevolo, C., Dameri, R.P., D'Auria, B.: Smart mobility in smart city. In T. Torre et al. (Eds.), Empowering Organizations (13–28). Springer (2016).
4. Bibri, S.E., Krogstie, J.: Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable Cities & Society **31**, 183–212 (2017).
5. Hyndman, R. J., Athanasopoulos, G.: Forecasting: principles and practice. 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3 (2021).
6. Metulini, R., 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).
7. Tettamanti, T., Varga, I.: Mobile phone location area based traffic flow estimation in urban road traffic. Columbia International Publishing, Advances in Civil and Environmental Engineering **1(1)**, 1-15 (2014).

CP decomposition of 4th-order tensors of compositions

Decomposizione CP di tensori composizionali di ordine 4

Violetta Simonacci, Tullio Menini and Michele Gallo

Abstract Multifold data structures are generally stored in high-dimensional objects defined as *n*th-order tensors. Generalization of trilinear decompositions such as the CANDECOMP/PARAFAC model can be used for modelling 4th order tensors. The application of these techniques is, however, quite limited due to procedural complexity and interpretational issues. These concerns increase when tensors contain data with a compositional structure. This work aims at addressing these difficulties through an application on Italian university staff.

Abstract *Strutture di dati complesse sono generalmente memorizzate in oggetti multidimensionali definiti come tensori di ordine n. Per modellare tensori di ordine 4, è possibile utilizzare generalizzazioni delle decomposizioni trilineari come il modello CANDECOMP/PARAFAC. L'applicazione di queste tecniche è, tuttavia, piuttosto limitata a causa della loro complessità procedurale e interpretativa. Tali difficoltà aumentano poi nel caso in cui i tensori contengano dati con una struttura composizionale. Questo lavoro mira ad affrontare tali problemi attraverso un'applicazione sul personale universitario italiano.*

Key words: CoDa, CANDECOMP/PARAFAC, logratio, higher order decomposition, parameter estimation

Violetta Simonacci
Dept. of Social Science, University of Napels Federico II, Naples, Italy
e-mail: violetta.simonacci@unina.it

Tullio Menini
Dept. of Human and Social Sciences, University of Naples - "L'Orientale", Naples, Italy
e-mail: menini@unior.it

Michele Gallo
Dept. of Human and Social Sciences, University of Naples - "L'Orientale", Naples, Italy
e-mail: mgallo@unior.it

1 Introduction

Complex social phenomena are the results of different layers of information continuously interacting at repeated occasions. As data-storing capabilities become virtually unbounded, finding effective ways of modeling together multiple entities has become an ongoing challenge.

Tensors are the preferred algebraic architecture for storing complex data and describing multilinear relationships between entities in a compact form. A generic *n*th-order tensor stores data along *n* indices and can be described as a generalization of simple structures such as scalars, vectors and matrices which are special cases of 0-order (no index), 1st-order (1 index) and 2nd-order (2 indices) tensors.

Tensor data structure may presents additional challenges besides a multidimensional variability structure. Let us think of tensor with proportion values (e.g. percentages, shares, parts of a total), defined in statistical literature data as Compositional Data (CoDa). Such data are characterized by a biased covariance structure which can be modeled only in relative terms [1] and requires special tools.

Tensor decompositions techniques can come quite handy when dealing with multilinear data. These tools allow capturing the multidimensional information in a tensor by breaking it down in sets of simpler objects, generally lower order tensors. The two most commonly used techniques for the decomposition of *n*th-order tensors are the Higher-Order TUCKER and CANDECOMP/PARAFAC (CP) models [9, 5]. The TUCKER model is more suitable for summarizing large information into condensed sets of variables, thus, it is the preferred method for tensor compression and variability structure descriptions. The CP method is more appealing when trying to retrieve a meaningful underlying structure. This is because this model provides a unique solution under mild conditions [8].

The higher order CP model can be easily adapted to compositional data by use of log-ratio transformations which, applied prior to the decomposition, do not alter its procedural steps but call for an additional interpretability effort.

Multilinear decomposition for tensors of order higher than 3 are occasionally used in Chemistry related fields, however, their applications in social sciences is uncommon. This is mainly due to model complexity which makes these tools unfriendly for non-experts. For tensors of compositions the degree of complexity increases even more, thus, compositional adaptations of n-th order decompositions are completely absent in social sciences.

Given these considerations the aim of this work is to address two issues which cause the infrequent use of these tools, namely, parameter estimation ambiguities and interpretability concerns. The focus will be only on the CP procedure because its desirable uniqueness makes it more vulnerable to efficiency and algorithmic problems.

In order to reach this goal, an application on University teaching staff in Italy recorded by macro-region, disciplinary field, role and year will be presented. Specifically, a 4th-order tensor is considered in which disciplinary field shares are treated as compositional data. After following CoDa methodology by extending the strategy proposed for tridimensional arrays to a 4-way tensor, the CP model will be

computed. Results will be analyzed by paying careful attention to the advantages of using such procedure and to the estimating problems of current algorithms in a compositional setting.

In Section 2 tensor notation is explained and the dataset is briefly introduced; in Section 3 the methodology is outlined for the four-way CoDa-CP procedure and in Section 4 some initial consideration are conveyed.

2 Tensor notation a data

Let us consider a 4th-order tensor $\mathscr{T}$ with data arranged over the four indices $[1,\ldots,i,\ldots,I]$, $[1,\ldots,j,\ldots,J]$, $[1,\ldots,k,\ldots,K]$ and $[1,\ldots,l,\ldots,L]$. Its generic element is denoted by t_{ijkl}. The information contained in such tensor can be rearranged in many ways to focus on index relationships. The simplest way is to consider its composing vectors, generally referred to fibers. There are four types of fibers, one for each index so that I-,J-, K- and L-dimensional vectors can be identified as a generalization of rows and columns of a matrix. It is clear that there are as many fibers of a type as the product of the remaining indices, e.g. there are IKL fibers or rows $\mathbf{t}_{i:kl}$ with dimension J.

The tensor $\mathscr{T}$ can also be rearranged in 3rd-order blocks obtained by combining two of the four modes together into pseudo-fully stretched arrays $\underline{\mathbf{T}}_I(I \times JK \times L)$, $\underline{\mathbf{T}}_J(J \times KL \times I)$, $\underline{\mathbf{T}}_K(K \times LI \times J)$ and $\underline{\mathbf{T}}_L(L \times IJ \times K)$ [6].

Each of these tridimensional blocks can be seen as a set of slices, namely 2nd-order sections obtained by fixing one the three indices of the pseudo-fully stretched arrays and varying the remaining two. Specifically, it is possible to identify four sets of frontal slices $\mathbf{T}_{::l}(I \times JK)$, $\mathbf{T}_{::i}(J \times KL)$, $\mathbf{T}_{::j}(K \times LI)$ and $\mathbf{T}_{::k}(L \times IJ)$.

These alternative notations are only some of the many ways tensor information can be rearranged presented here to aid methodological explanations.

A 4th-order tensor presents a compositional structure if the elements of at least one of the fiber types describe the parts of a whole. Following conventions, let us assume that the J-dimensional fibers or rows are CoDa. Formally we have that the generic row $\mathbf{t}_{i:kl}$ is a compositional vector if it describes a point bounded in a subspace of $\Re^J_+$ known as simplex and defined as:

$$S^J = \left\{ \left(t_{i1kl_{(1)}}, \ldots, t_{iJkl} \right) : t_{i1kl} \geq 0, \ldots, t_{iJkl} \geq 0; t_{i1kl} + \ldots + t_{iJkl} = \kappa \right\} \tag{1}$$

where κ is a positive constant. To operate within this subspace special operations and rules known as Aitchison geometry must be followed. Alternatively CoDa vectors can be conveyed in real space coordinates by transforming them into log-ratios. Several transformations have been proposed in the literature, however, for brevity purposes only centered log-ratio (*clr*) coordinates are introduced. This function generates an isometric mapping between S^J and a hyperplane of $\Re^J$ in this fashion:

$$\mathbf{z}_{i:kl} = \text{clr}(\mathbf{t}_{i:kl}) = \left[\log\frac{t_{i1kl}}{\text{g}(\mathbf{t}_{i:kl})}, \ldots, \log\frac{t_{ijkl}}{\text{g}(\mathbf{t}_{i:kl})}, \ldots, \log\frac{t_{iJkl}}{\text{g}(\mathbf{t}_{i:kl})}\right] \quad \text{with } \text{g}(\mathbf{t}_{i:kl}) = \sqrt[J]{\prod_{j=1}^{J} t_{ijkl}} \tag{2}$$

These coordinates have the limit of yielding a pure multicollinear structure, which may cause estimating issues. As demonstrated in [2, 3] for 3rd-order tensors, *clr*-coordinates can be directly modeled with standard statistical tools. For the 4th-order tensor $\mathscr{T}$ a four-way CP model can be implemented, than results are translated back into compositional terms.

After clarifying tensor notation, the application of interest can be described in these terms. The dataset contains information on University teaching staff in Italy arranged over 4 directions with the following dimensions: 5 macro-region, 14 disciplinary fields, 3 role and 5 year, yielding a small tensor $\mathscr{T}$ with dimensions $(I = 5 \times J = 14 \times K = 3 \times L = 5)$.

For each macro-region, the partitioning among different disciplinary fields of the total number employee can be described as a compositional problem. Each row vector can thus be transformed as shown in eq.2 obtaining a new 4th-order tensor $\mathscr{Z} \in \mathbb{R}^{5\times14\times3\times5}$. This tensor can be decomposed with the CoDa-CP model as showed in the following section.

3 Four-way CoDa-CP model

Four-way CoDa-CP is an estimating model based on the polyadic decomposition which aims at providing the best low rank approximation of the tensor $\mathscr{Z} = \hat{\mathscr{Z}} + \mathscr{E}$, where $\mathscr{E}$ is the tensor of residuals. Here, the tensor is decomposed into the sum of a finite $f = 1, \ldots, F$ number of 1st-order factors $\mathbf{a}_f$, $\mathbf{b}_f$, $\mathbf{c}_f$ and $\mathbf{d}_f$:

$$\hat{\mathscr{Z}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{d}_f \tag{3}$$

The F terms of this decomposition can be arranged in four factor matrices $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_f, \ldots, \mathbf{a}_F]$, $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_f, \ldots, \mathbf{b}_F]$, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_f, \ldots, \mathbf{c}_F]$ and $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_f, \ldots, \mathbf{d}_F]$.

The model can also be rewritten using the pseudo fully starched array slice notation as follows:

$$\mathbf{Z}_{::l} = \mathbf{A}\text{diag}(\mathbf{d}_{(l)})(\mathbf{C} \odot \mathbf{B})^t + \mathbf{E}_{::l} \quad l = 1, \cdots, L \tag{4}$$

$$\mathbf{Z}_{::i} = \mathbf{B}\text{diag}(\mathbf{a}_{(i)})(\mathbf{D} \odot \mathbf{C})^t + \mathbf{E}_{::i} \quad i = 1, \cdots, I \tag{5}$$

$$\mathbf{Z}_{::j} = \mathbf{C}\text{diag}(\mathbf{b}_{(j)})(\mathbf{A} \odot \mathbf{D})^t + \mathbf{E}_{::j} \quad j = 1, \cdots, J \tag{6}$$

$$\mathbf{Z}_{::k} = \mathbf{D}\text{diag}(\mathbf{c}_{(k)})(\mathbf{B} \odot \mathbf{A})^t + \mathbf{E}_{::k} \quad k = 1, \cdots, K \tag{7}$$

Here $\odot$ is the Khatri-Rao product and $\text{diag}(\mathbf{d}_{(l)})$, $\text{diag}(\mathbf{a}_{(i)})$, $\text{diag}(\mathbf{b}_{(j)})$ and $\text{diag}(\mathbf{c}_{(k)})$ denote the diagonal matrices extracting the *l*th, *i*th, *j*th and *k*th rows of the factor matrices respectively.

The four-way CoDa-CP model is unique under mild conditions and is generally estimated through a least-squares loss function. Estimation problems may, however, occur, such as solution degeneracies [10] and slow convergence, especially for collinear data [7].

4 Preliminary considerations

One of the best ways to unveil the latent structure of 4th-order tensor is to carry out a CP decomposition. The uniqueness of the CP model makes this procedure both appealing and harder to estimate with respect to other techniques for the decomposion of 4th-order tensors as the TUCKER model [9]. Many difficulties may arise when estimating CP parameters connected to both efficiency and accuracy of the solution. Multicollinearity, typical of *clr*-coordinates, makes this issues even more pressing.

Several procedure have been proposed over the years to cope with these difficulties, all with different points of strenght and fallacies. The problem, however, is generally dealt with for the simpler case of 3rd-order tensors.

In this work, by considering the 4th-order tensor of University teaching staff data we are going to tackle two challenges: 1) show the potential of the four-way CoDa CP methodology with respect to other, more common, modeling tools; 2) explore the estimation problem of the CP model in the generalized framework of 4-way compositional data by extending the work of [4].

References

1. Aitchison, J.: The statistical analysis of compositional data. Chapman & Hall (1986)
2. Gallo, M.: Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data. In: Advanced dynamic modeling of economic and social systems, pp. 209-221, Springer (2013)
3. Gallo, M., Simonacci, V.: A procedure for the three-mode analysis of compositions. Electronic Journal of Applied Statistical Analysis **6**(2), 202–210 (2013)
4. Gallo, M., Simonacci, V., Di Palma, M.A.: An integrated algorithm for three-way compositional data. Quality & Quantity **53**(5), 2353–2370 (2019)
5. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an "explantory" multi-modal factor analysis. UCLA working papers in phonetics **16**, 1–84 (1970)
6. Kang, C., Wu, H.L., Yu, Y.J., Liu, Y.J., Zhang, S.R., Zhang, X.H., Yu, R.Q.: An alternative quadrilinear decomposition algorithm for four-way calibration with application to analysis of four-way fluorescence excitation–emission–ph data array. Analytica chimica acta **758**, 45–57 (2013)
7. Kiers, H.A.: A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. Journal of Chemometrics **12**(3), 155–171 (1998)
8. Sidiropoulos, N.D., Bro, R.: On the uniqueness of multilinear decomposition of n-way arrays. Journal of chemometrics **14**(3), 229–239 (2000)

9. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
10. Zijlstra, B.J., Kiers, H.A.: Degenerate solutions obtained from several variants of factor analysis. Journal of Chemometrics: A Journal of the Chemometrics Society **16**(11), 596–605 (2002)

A strategy of analysis of symmetry and skew-symmetry in asymmetric relationships

Una strategia di analisi di simmetria ed emi-simmetria in relazioni asimmetriche

Giuseppe Bove

Abstract Proximity matrices are frequently asymmetric and analysed by the additive decomposition in the symmetric and the skew-symmetric component. Models for joint or separate graphical representations of the two components have been proposed by many authors. A strategy for such graphical analysis is proposed and applied to a data set regarding relationships between members of an organizational structure of a Japanese firm.

Abstract *Le matrici di prossimità sono spesso asimmetriche ed analizzate attraverso la decomposizione additiva nelle componenti simmetrica ed emisimmetrica. Molti autori hanno proposto modelli per la rappresentazione grafica congiunta o separata delle due componenti. In questo lavoro si propone una strategia per tale analisi grafica, applicandola a dati riguardanti la struttura organizzativa di una impresa giapponese.*

Key words: proximity data, asymmetry, graphical representation

1 Introduction

In many disciplines such as economics, sociology, marketing research and other behavioral sciences, asymmetric proximities between pairs of entities in a set (e.g., import-export data, sociomatrices, brand switching, flows and migration data, etc.) are frequently studied to detect meaningful relationships. In particular, members in organizational structures (e.g., employees in a firm, students at school, friends on web,

[1] Giuseppe Bove, Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre; email: giuseppe.bove@uniroma3.it

etc.) are usually related in an asymmetric way (e.g., 'A approaches B for help and advice', 'A is a friend of B', 'A sends an e-mail to B', etc.). From now on, a proximity is denoted by the symbol ω, and the values on the pairs (i,j) $(i,j = 1,2,\dots,n)$ collected in a $(n \times n)$ data matrix $\mathbf{\Omega} = (\omega_{ij})$. The unique additive decomposition $\mathbf{\Omega} = \mathbf{M} + \mathbf{N}$, with matrix $\mathbf{M} = (m_{ij} = 1/2(\omega_{ij} + \omega_{ji}))$ symmetric and matrix $\mathbf{N} = (n_{ij} = 1/2(\omega_{ij} - \omega_{ji}))$ skew-symmetric ($\mathbf{N} = -\mathbf{N}'$) has been largely studied and applied. In this contribution a strategy of analysis will be presented to represent jointly or separately in diagrams the symmetric and the skew-symmetric components, in order to detect easily the information regarding symmetry and asymmetry. The different steps of the strategy will be presented by using a data matrix regarding the organizational structure of a Japanese firm.

2 A first look at symmetry and skew-symmetry

Krackhardt (1987) reported an empirical study concerning relationships among 21 managers of a small manufacturing organization, producing high-tech machinery. The kind of relationship to be analysed (of the type 'A approaches B for help and advice') was submitted to all 21 managers, who were presented with a questionnaire of 21 questions (one for each manager). Therefore, the original data consist of 21 square (21×21) matrices (termed slices) of dichotomous values 0/1 (available in Krackhardt, 1987, Appendix A, p. 129), each provided by one manager. The aggregated proximity matrix resulting by summing all the slices is provided in Okada (2011, Table 2). Each off-diagonal entry (i,j) represents the number of managers who responded that manager i goes to manager j for help or advice at work. Diagonal entries are not defined. The proximity is assumed as a measure of closeness (or similarity) between managers. The management consists of one president (label 7), four vice presidents (labels 2, 14, 18, 21) and sixteen supervisors. Each vice president heads up a department, and each of the sixteen supervisors belongs to one of the four departments (see Okada, 2011, Table 1). As often happens, the contribution of the two matrices $\mathbf{M}$ and $\mathbf{N}$ (not reported here) to the decomposition of matrix $\mathbf{\Omega}$ is very unbalanced, the two sum of squares ratios are 97% and 3%, respectively. The correlation coefficient between $\mathbf{\Omega}$ and its transpose $\mathbf{\Omega}'$ is 0.64 that confirms the dominance of the symmetry in the observed data matrix (the correlation is 1 when $\mathbf{\Omega}$ is symmetric and -1 when it is skew-symmetric). However, the skew-symmetric component can reveal interesting aspects of the relationship 'A approaches B for help and advice', in spite of its small contribution. In the following, the two components will be analysed by different spatial models that allow to detect easily in diagrams the relationships for help and advice between managers at work, and the centrality role of the president and vice-presidents.

A method for displaying simultaneously the symmetric and the skew-symmetric component based on the *drift vector model* is presented in Borg & Groenen (2005, p. 502). Firstly, a map for the symmetric component matrix $\mathbf{M}$ is provided by a symmetric MDS method, then the skew-symmetric information in $\mathbf{N}$ is incorporated

in the map by drawing arrows from each point *i* to any other point *j* of the map, so that lengths and directions correspond to the entries in the rows of the skew-symmetric matrix. When the number of rows (managers) of the proximity matrix is large, reporting all arrows can provide cluttered pictures, and it can be convenient to draw only the resultant (average) of the arrow bundle attached to each point. Even if the possibility to depict pair-wise skew-symmetries is lost, the average arrow (called *drift vector*) can allow to better detect how the symmetric and the skew-symmetric data components are related to each other.

In Figure 1 the drift vectors for the 21 managers are shown, points labelled by numbers represent managers. Distances between points represent entries of matrix **M** (*Stress*-I=0.26), the higher m_{ij} the smaller the distance. Departments are represented by points labelled D1, D2, D3, D4, obtained as the averages of the coordinates of the corresponding managers. First, we remark that the president (point 7) and three vice-presidents (points 2, 14 and 18) are positioned in the centre of the diagram, that means they are the main destinations of the requests for advice. The distance between managers reflects quite well the departments they belong to, with just a few exceptions (managers 3 and 8). The vice president labelled 21 is located in the fourth quadrant, close to the managers of Department 1 that he heads up (this suggests that his advice is almost exclusively oriented to the managers of his Department). The pattern defined by the drift vectors show that the asymmetries are not random. Most of the arrows attached to the supervisors have a direction opposite to the centre of the diagram, confirming the central role and the dominance of the president and the vice presidents. The lengths of the drift vectors show that asymmetries are mainly concentrated in Department 2. However, a detailed analysis of matrix **N** is possible by the explicit models for skew-symmetry presented in the next sections.

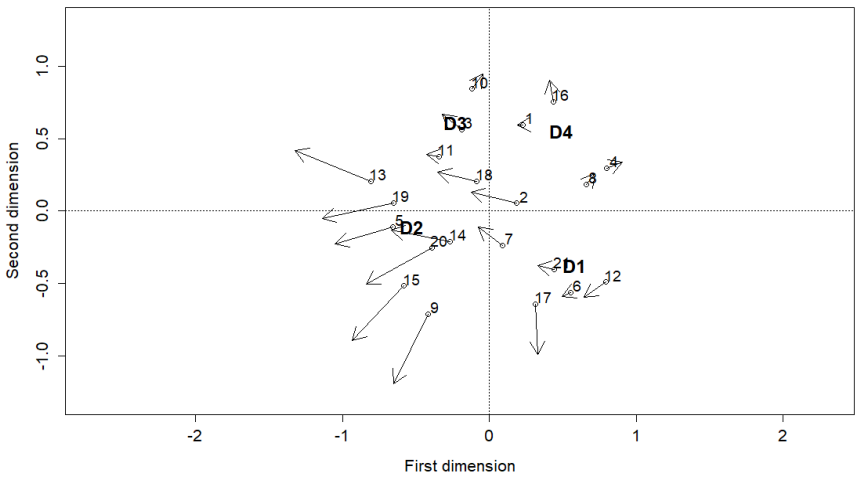


Fig. 1 Representation of symmetry (distances) and skew-symmetry (drift vectors)

3 Models for one-dimensional skew-symmetry

A simple model to represent skew-symmetry in one dimension is $n_{ij} = (r_i - r_j) + \varepsilon_{ij}$, ($i$=1,…,$n$). A unique least-square solution cannot be identified, because, given the solutions $\hat{r}_i$, any translation $\hat{r}_i + c$ by any constant c represents equivalent solution. Without loss of generality, if the constraint $\sum_{i=1}^{n} r_i = 0$ is assumed, it is easy to prove that least-squares estimates of r_i (i=1,…,n) are the averages of the n_{ij} values within each row i over the columns, that is $\hat{r}_i = \frac{1}{n}\sum_{j=1}^{n} n_{ij}$. The variance accounted for by the one-dimensional model is 65%. The estimates $\hat{r}_i$ can be represented separately on a straight line (Figure 2) or they can be attached as radii on the points of Figure 1 (Figure 3) by the *radius-distance model* (Okada & Imaizumi, 1987). The radii of the circles can be obtained from the nonnegative estimates $\hat{r}_i$ by an appropriate translation by a constant c such that $min(\hat{r}_i + c) = 0$ (the smallest radius is equal to zero, radii can be also rescaled to be easily comparable).

In Figure 2, the skew-symmetries can be analysed as differences between the estimates $\hat{r}_i$ (oriented or signed Euclidean distances). The president and the vice presidents are positioned on the left side of the axis, so most of the skew-symmetries with the supervisors are negatives, that means supervisors ask for help and advice to the president and vice presidents more frequently than the reverse. The order from the left to the right side of the axis reflects the dominance order between the managers.

In Figure 3, when the circle around point i is larger than circle around point j, then the estimate of the skew-symmetry n_{ij} is positive and the estimate of the skew-symmetry n_{ji} is negative. So, the president and the vice presidents have the smallest radii because they are characterized by negative skew-symmetries (vice president 2 has radius zero).

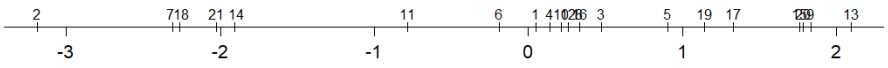


Fig. 2 Representation of skew-symmetry on one dimension (manager labels above the line)

4 Singular Value Decomposition of skew-symmetry

Since **M** and **N** are orthogonal (i.e., *tr*(**MN**) = 0), Constantine and Gower (1978) suggested to representing the two components separately.

A strategy of analysis of symmetry and skew-symmetry

An advantage of this approach is that we can represent skew-symmetry in more than one dimension to obtain a better approximation of matrix **N**. A representation in a plane (named *Gower diagram*) is obtained by the singular value decomposition of **N** (Figure 4). The variance of **N** accounted for in the diagram is 74%. The interpretation of the diagram in Figure 4 is not in terms of distances (or inner products) but in terms of areas (Gower, 1977).

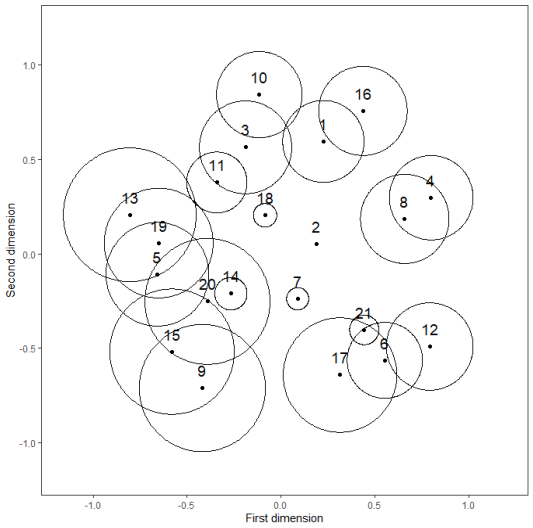


Fig. 3 Representation of proximities by the radius-distance model

So, for instance, the area of the triangle depicted in the diagram approximates the skew-symmetry between the president and the vice president 2 and it is positive counterclockwise, that means vice president 2 ask for help and advice to the president more frequently than the reverse (this aspect was incorrectly represented in one dimension, Figures 2 and 3). Most of the triangles from supervisors to the president and the vice presidents have large areas and the skew-symmetries are positive, that confirm that supervisors ask for help and advice to the president and the vice presidents more frequently than the reverse.

Other approaches to the analysis of symmetry and skew-symmetry in proximity data can be found in Bove et al. (2021, Chapter 3).

5 Conclusion

A strategy for the graphical analysis of the symmetric and the skew-symmetric components of asymmetric proximity matrices has been described. Models are chosen following their levels of complexity (i.e., the number of parameters), firstly applying

parsimonious models with joint graphical representations of the two components easy to interpret. A separate analysis of symmetry and skew-symmetry should be preferred when an adequate level of data approximation is not obtainable by models for joint representation.

Information concerning some R functions for fitting the models presented can be found in Bove et al. (2021).

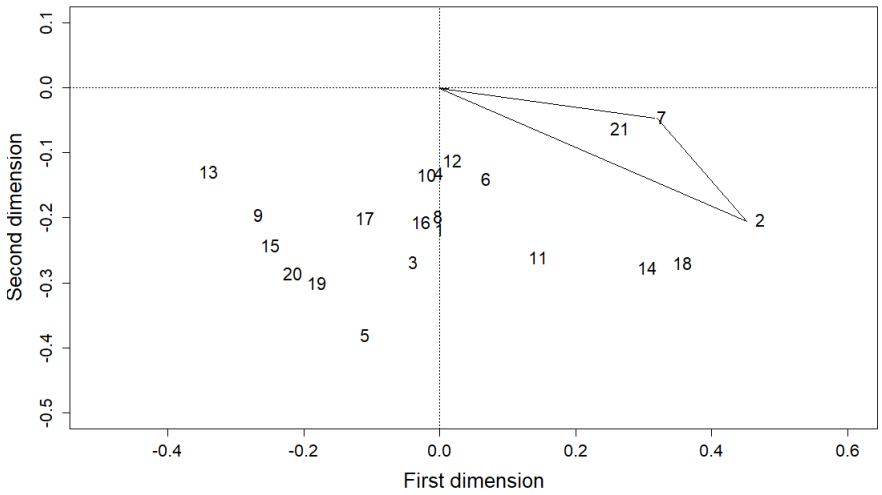


Fig. 4 Representation of skew-symmetry by the singular value decomposition

References

1. Borg I., Groenen, P.J.F.: Modern Multidimensional Scaling. Theory and Applications. (Second Edition). Springer, New York (2005)
2. Bove, G., Okada, A., Vicari, D.: Methods for the analysis of asymmetric proximity data. Springer Nature, Singapore (2021)
3. Constantine, A.G., Gower, J.C.: Graphical representation of asymmetric matrices. Applied Statistics **3**, 297-304 (1978)
4. Gower, J.C.: The analysis of asymmetry and orthogonality. In: Barra J.R. et al. (eds), Recent Developments in Statistics. North Holland, Amsterdam, 109-123 (1977)
5. Krackhardt, D.: Cognitive social structures. Social Networks **9**, 109–134 (1987)
6. Okada, A.: Centrality of asymmetric social network: singular value decomposition, conjoint measurement, and asymmetric multidimensional scaling. In: Ingrassia S. et al. (eds.), New perspectives in statistical modeling and data analysis. Springer, Heidelberg, 219–227 (2011)
7. Okada, A., Imaizumi, T.: Non-metric multidimensional scaling of asymmetric similarities. Behaviormetrika **21**, 81-96 (1987)

Session of free contributes SCL2 *–Industry and Society*
Chair: Germana Scepi

Toward an early detection of SME's default with websites' indicators

Verso un rilevamento precoce del fallimento delle PMI tramite i loro siti web

Lisa Crosato, Josep Domenech, Caterina Liberati

Abstract Small and medium enterprises (SMEs) contribution to the European Union economy has always been relevant, for both value added and the creation of jobs. On the same time, SMEs are more fragile and likely to default as compared to larger firms so to induce public policies targeted on them. Specific default prediction models, accounting for SMEs idyiosincratic traits, are based on several types of data, mainly accounting indicators. We explore the possibility of complementing accounting information with data scraped from the firms' websites.

Abstract *Le piccole e medie imprese (PMI) costituiscono l'ossatura dell'economia dell'Unione Europea, con la formazione di gran parte del valore aggiunto e la creazione di posti di lavoro. Purtroppo le PMI sono allo stesso tempo pi fragili e a rischio di fallimento rispetto alle imprese di dimensioni maggiori. Anche per questo motivo, in tutti i paesi dell'Unione Europea, esistono dei programmi di sostegno specificatamente indirizzati alle PMI. I modelli di previsione del fallimento per questa categoria di imprese pi utilizzati analizzano dati derivati dai bilanci d'impresa. La nostra proposta di aggiungere a questi ultimi anche le caratteristiche dei siti web delle imprese.*

Key words: SMEs default, web scraping, kernel discriminant analysis

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

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

Caterina Liberati
DEMS, University of Milano-Bicocca
e-mail: caterina.liberati@unimib.it

1 Introduction

Preventing SMEs default, financing most promising firms and sustaining them in difficult times [4] means protecting 99% of all enterprises in the EU (Eurostat), as well as the largest part of the European value added and jobs (56.4% and 66.6% respectively, [12]). Thus, it is not by chance that single governments and European Institutions promote SMEs-addressed support policies [10, 9, 11].

Accordingly, there is a vast literature studying SMEs default factors in European countries [6, 13, 25], mainly on the basis of accounting indicators derived from balance sheets or rarely on other kinds of data [7, 26, 28]. All of these data sources suffer of a large delay between their availability and their reference period.

In this paper we explore the possible use of websites as a relevant source of information for an earlier detection of SMEs default [8] to be of help for avoiding both credit and public funding misallocation. On the one hand, web content data clearly require substantial efforts in data retrievement, selection, cleaning and ultimately analysis with respect to traditional sources of data. On the other hand, website information is free, assures finer granularity, a larger coverage of the firms' population and, most importantly, up-to-dateness. Previous works in the literature have used corporate websites to retrieve online proxies of firms' economic characteristics, such as corporate culture [24], firm performance [22], firm strategies [19] or innovation [2].

The online indicators can be generated after manually reviewing firm websites, or automatizing the process via web crawling and scraping techniques, as in [5]. The latter technique allows for a new approach to systematically monitor companies based on the changes (or updates) of these company websites. The study can be done retrospectively with the Wayback Machine of the Internet Archive, which is a digital library storing Internet sites and their evolution in time.

Using about 700 spanish SMEs sampled from the SABI -Sistema de Análisis de Balances Ibéricos (Bureau van Dijk)- we intend to build up a unique dataset combining the accounting (offline) indicators with our new online indicators. Our first purpose is to classify and describe the characteristics of a website which can be used to discriminate between surviving and defaulted firms. The joint use of online and offline information for enhancing correct prediction of default will be explored through nonlinear discriminant analysis keeping logistic regression as a benchmark.

2 Website features

In simple terms, websites are a set of documents (generally HTML) which are stored in a web server. HTML is a mark-up language which describes the content of a web page, providing the page text and some semantics on how to interpret it and its structure.

Business websites can be analyzed (or mined) from two different perspectives[1]: web structure mining and web content mining. While the former focuses on how the different pages are linked, the latter concentrates on understanding the semantics and the meaning of the contents. This section focuses only on content mining, since it is the closest approach to the business activity. Content mining has been approached below from two different perspectives: i) the mining of the text composing the business website, and ii) the mining of the HTML code describing the text.

2.1 Textual Content

Most business websites include a significant part of text, which usually describes the main activity of the company. Therefore, it is expected that information such as the sector in which it operates and its market orientation (e.g., national or international, final consumer or other businesses) emerges from the analysis of this text. Changes in this text would also mean that the company is changing its behavior to some extent, and thus, it is still alive and investing in the website.

The feature extraction from the website text can be done by means of general text mining techniques, which encompass a wide variety of the processes for discovering information in textual data. The particular techniques that are used in this paper are stemming and the bag-of-words model. Stemming consists on reducing word found to its stem (e.g., *industry* and *industrial* are reduced to *industri*). The bag-of-words model is a representation of the frequency of occurrence of each word in a text. This way, the text is converted into a set of variables that can be later used by the classifiers.

2.2 HTML code

The HTML language is a standard defined by the World Wide Web Consortium that describes the elements of a web page by using tags. These tags are useful to describe the interaction (e.g., defining hyperlinks or forms), appearance (e.g., bold or italics), and the structure (e.g., defining lists or different blocks) of a web page. Since it has evolved through years, the tags used in a website may be related to how updated it is. How these tags are used provide relevant information to capture the underlying behavior of companies. For instance, FORM is usually employed to interact with the company/site. EMBED is generally employed to include Flash technology, which is currently being abandoned. LAYER is another legacy tag, used decades ago to design the web page (browsers support it for compatibility reasons). Hyperlinks are defined with A tags, whose analysis allows us to detect connections with external agents and website structure. They can be further analyzed by checking which text is

[1] There is a third perspective (web usage mining) that can only be used by the owners of the website, which is not the case for the research presented in this paper.

included in the hyperlink reference, or href (e.g., twitter, government, associations), or which file extensions are used, as they are related to the underlying technology (php/asp/htm...) or to which information is offered (pdf, xls...). The number of images (IMG tags) in the website or the extension used (png, jpg...) is related to depth and quality of the information provided. Some other tags, such as META and LINK, are also related to the technology used in the company website.

3 Research design and Methodology

The research design of our study presents two challenges to be tackled: the dimensionality of the data matrix and the modelling of non linear patterns. The first task was to select the variables relevant for our analysis among the -more than 15.000- total websites' features. We work with binary indicators simply obtained as presence or absence of a given feature in a firm's website. In order to proceed to a first screening, we have selected as relevant features those with higher occurrences among 100 runs of a LASSO regression. The final subset of variables counts 50 binary indicators that were transformed into 41 numerical orthogonal factors, via Multiple Correspondence Analysis [16], to allow the application of any classifier.

As about the second task, we referred to the financial literature, where the bankruptcy prediction of the SMEs has been widely studied using different statistical techniques. Since this classification problem is hard to solve, also due to the overbalance between survived and defaulted companies, it requires the application of machine learning models that generally outperform the linear ones. Indeed, the majority of studies in this field have successfully employed non linear models as Deep Learning [20], Boosting [18] Neural Networks [27, 3] keeping the z-score proposed by [1] or the logistic regression [17] as a benchmark. In our work we use Kernel Discriminant Analysis (KDA) [23] since kernel-based algorithms have demonstrated their predictive power in screening SMEs [15, 29, 8].

More in detail, given a training set $\mathscr{I}_{XY} = \{(x_1, y_1), ..., (x_n, y_n)\}$ of size n where data $x_i \in R^p$ and labels $y_i \in \{1, 1\}$, the objective of KDA is to devise a decision function $f(x)$ as a combination of features that separates two classes of objects. Due to the recurrent non-linear separability of training data in the input space R^p, in most cases one associates the training data x_i to some feature space $\mathscr{F}$ through a non-linear mapping

$$\phi : x_i \mapsto \phi(x_i) \tag{1}$$

We can refer to $\mathscr{F}$ as a Reproducing Kernel Hilbert Space (RKHS) if the Mercers theorem is satisfied [21].

As in the Fisher Discriminant Analysis [14], maximizing the ratio of Between and the Within covariance matrices (computed in the Feature Space) is then used to define a separating hyperplane in $\mathscr{F}$ with direction vector w

$$g(x) = w^T \phi(x) + b \tag{2}$$

where w is the weight vector in RKHS, and $b \in R$ is the bias term.

Optimal generalization of kernel-based method still depends on the selection of a suitable kernel function out of the many maps proposed in the literature. Among the common kernel already employed, we have selected the Cauchy, Laplace, Multiquadric and RBF functions due to their remarkable performances.

References

1. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance **23**(4), 589–609 (1968)
2. Axenbeck, J., Breithaupt, P. Innovation indicators based on firm websites - which website characteristics predict firm-level innovation activity? PloS one **16**(4), e0249,583 (2021)
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**(6), 627–635 (2003)
4. Belghitar, Y., Moro, A., Radić, N. When the rainy day is the worst hurricane ever: the effects of governmental policies on smes during covid-19. Small Business Economics pp. 1–19 (2021)
5. Blazquez, D., Domenech, J. Web data mining for monitoring business export orientation. Technological and Economic Development of Economy **24**(2), 406–428 (2018)
6. Ciampi, F. Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of Italian firms. Journal of Business Research **68**(5), 1012–1025 (2015)
7. Cornée, S. The relevance of soft information for predicting small business credit default: Evidence from a social bank. J. Small Bus. Manag. **57**(3), 699–719 (2019)
8. Crosato, L., Domenech, J., Liberati, C. Predicting SMEs default: Are their websites informative? Economics Letters **204**, 109,888 (2021)
9. Cultrera, L., et al. Evaluation of bankruptcy prevention tools: evidences from COSME programme. Econ. Bull. **40**(2), 978–988 (2020)
10. Dvouletỳ, O., Srhoj, S., Pantea, S. Public SME grants and firm performance in european union: A systematic review of empirical evidence. Small Bus. Econ. pp. 1–21 (2020)
11. Dvouletỳ, O., Srhoj, S., Pantea, S. Public sme grants and firm performance in european union: A systematic review of empirical evidence. Small Business Economics **57**(1), 243–263 (2021)
12. European Commission. Annual Report on European SMEs 2018/2019. Tech. rep. (2019)

13. Fantazzini, D., Figini, S. Default forecasting for Small-Medium Enterprises: Does heterogeneity matter? International Journal of Risk Assessment and Management **11**(1-2), 138–163 (2009)
14. Fisher, R.A. The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
15. Gordini, N. A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Systems with Applications **41**(14), 6433–6445 (2014)
16. Greenacre, M.J. Theory and applications of correspondence analysis (1984)
17. Hosmer, D., Lemeshow, S. Applied Logistic Regression. Wiley (2000)
18. Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., Kou, S. Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decision Support Systems **140**, 113,429 (2021)
19. Llopis, J., Gonzalez, R., Gasco, J. Web pages as a tool for a strategic description of the spanish largest firms. Information processing & management **46**(3), 320–330 (2010)
20. Mai, F., Tian, S., Lee, C., Ma, L. Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research **274**(2), 743–758 (2019)
21. Mercer, J. Functions of positive and negative type, and their connection the theory of integral equations. Philosophical transactions of the royal society of London, A **209**, 415–446 (1909)
22. Merono-Cerdan, A.L., Soto-Acosta, P. External web content and its influence on organizational performance. European Journal of Information Systems **16**(1), 66–80 (2007)
23. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Muüller, K.R. Fisher discriminant analysis with kernels. In: Neural networks for signal processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop, pp. 41–48 (1999)
24. Overbeeke, M., Snizek, W.E. Web sites and corporate culture: A research note. Business & Society **44**(3), 346–356 (2005)
25. Succurro, M., Mannarino, L., et al. The Impact of Financial Structure on Firms' Probability of Bankruptcy: A Comparison across Western Europe Convergence Regions. Regional and Sectoral Economic Studies **14**(1), 81–94 (2014)
26. Tobback, E., Bellotti, T., Moeyersoms, J., Stankova, M., Martens, D. Bankruptcy prediction for SMEs using relational data. Decis. Support Syst. **102**, 69–81 (2017)
27. West, D. Neural Network Credit Scoring Models. Computers & Operations Research **27**(11-12), 1131–1152 (2000)
28. Yin, C., Jiang, C., Jain, H.K., Wang, Z. Evaluating the credit risk of SMEs using legal judgments. Decision Support Systems **136**, 113,364 (2020)
29. Zhang, L., Hu, H., Zhang, D. A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. Financial Innovation **1**(1), 1–21 (2015)

The Culture and Values of Italians after two years of pandemic

La Cultura ed i Valori degli Italiani dopo due anni di pandemia

Raffaele Angelone

Abstract The study gives us a picture of Italians who are less traditionalist, less religious, and more concerned with self-care and self-affirmation. For most of them, freedom of thought and choice, chasing their dreams is more important than economic independence. Self-care is accompanied by a growing attention to the environment and a rediscovery of social responsibility and solidarity as fundamental components for improving the social environment in which we live.

.

Abstract Lo studio ci restituisce un'immagine degli Italiani meno tradizionalisti, religiosi, ma sempre più attenti a valori di libertà di scelta, alla cura e all'affermazione di sè. Per la maggioranza di loro la liberà di pensiero e di scelta e rincorrere i propri sogni è più importante dell'indipendenza eeconomica. La cura di sè si accompagna ad una crescente attenzione rispetto all'ambiente e ad una riscoperta della responsabilità sociale e della solidarietà come componenti fondamentali per migliorare l'ambiente sociale in cui si vive

Key words: Values, Italians, Pandemic

The Methodology

[1]Raffaele Angelone, Contract Professor at University of Milan "La Bicocca";
raffaele.angelone@unimib.it

The analysis is based on a survey run by Demoskopea Research Institute to understand Italians' value systems after two years of Covid-19 pandemic. A sample of 1014 individuals representative of the adult population (18+ years old) was interviewed at the end of Sept. 2021. The interview was administered through a 15-minutes CAWI questionnaire. The segmentation was performed using the k-means method.

The Italians' Values – Main Results

A different and transformed Italy emerges from the pandemic, starting with its own values. Italians seem to focus on self-direction, wellbeing, and self-care, but they assign a growing importance to the context around them, showing a high interest in environmental issues. It is an extremely practical and pragmatic form of self-care. The focus is on health and physical wellbeing which does not translate into a quest for greater religiosity. Money and career find less space in the Italian reshaping of values.

Table 1: The Italians' Values Ranking [1]

The Values	***Importance Avg. Score*** ***(1=not important; 10= extremely important)***
Freedom of action	8,7
Attention to the environment	8,4
Self-confidence	8,3
Health	8,3
Stable personal relationships	8,3
Happiness	8,3
Fun	7,7
Solidarity	7,7
Social responsibility	7,5
Thrift	7,2
Modesty	7,2
Excitement	6,8
Traditional gender roles	6,1
Status	5,8
Wealth	5,8
Faith	5,6
Power	4,9

Adherence to these values is common to all geographical areas of the country and age groups with some differences:

- The South differs in being more traditionalist and more attached to values such as status, wealth, and power.
- Young people tend to be less traditionalist, more hedonist and in search of status
- Mature people tend to be more traditionalist and religious, more attentive to solidarity and social responsibility values

The Italians' Segmentations by Values

The clustering exercise highlighted the presence of three groups of Italians who differ in their approach to values and life:

- **The "Achievers"**. They represent 40% of population. They share the values of self-care, self-direction, attention to environment, social responsibility subscribed by most of population. Their approach to life is more sensitive to happiness and hedonistic values such as fun and stimulating experience. They are quite traditionalist, and the most interested group in power, wealth and status

- **"The Mindless"** They represent 30% of population. They are the least interested in self- direction, social responsibility, caring for environment and solidarity values. Their approach to life is moderately traditionalist and attracted to power.

- **"The Unconventionals**". They represent 30% of population. As "the Achievers" they share the attention to themselves, to ecology and social environment they live. Their approch to life is not traditionalist and they are less drawn to fun, excitement, and religious faith. They are also the least sensitive to power status and wealth

Table 2: The Values by Segment [1,2]

The Values	***Total Population Importance Avg. Score*** *(1=not important; 10= extremely important)*	***Achievers*** *Index vs. Total*	***Mindless*** *Index vs. Total*	***Unconventional*** *Index vs. Total*
Freedom of action	8,7	107	85	106
Attention to the environment	8,4	108	83	107
Self-confidence	8,3	108	85	104
Health	8,3	110	85	102
Stable personal relationships	8,3	108	85	104
Happiness	8,3	109	85	104
Fun	7,7	110	85	102
Solidarity	7,7	110	84	103
Social responsibility	7,5	111	83	102
Thrift	7,2	110	89	98
Modesty	7,2	111	93	93
Excitement	6,8	113	88	94
Traditional gender roles	6,1	127	109	55
Status	5,8	118	98	78
Wealth	5,8	113	99	83
Faith	5,6	135	102	52
Power	4,9	119	106	69

References

1. Angelone, R., Fondrini, G.: Survey on the Italians' Values (2021)
2. Angelone R.: Ricerche di Marketing: strumenti e tecniche dalla teoria alla pratica aziendale. PKE (2021)

Thematic evolution of Academic Medical Centers' research: a focus on Italian public owned AOUs in metropolitan areas

L'evoluzione delle tematiche di ricerca nei Centri Medici Accademici: un focus sulle strutture sanitarie italiane pubbliche in aree metropolitane

Massimo Aria, Corrado Cuccurullo, Luca D'Aniello and Maria Spano

Abstract In recent years, there is an increasing recognition of the potential value of research evidence as one of the many factors considered by policymakers and practitioners. Even more, in the case of medical science, the analysis of research and its impact is indispensable, considering its implications for public health. By means of science mapping techniques, we provide a tool for the visualization of strategic positioning of different Italian public owned Academic Medical Centers in terms of their research positioning. Our proposal aims to provide a conceptual framework for policymakers involved in healthcare institutions, and at the same time, for the institutions themselves, to direct their research activities towards increasingly innovative scenarios that consider the general landscape of current research.

Abstract *Negli ultimi anni, c'è un crescente riconoscimento del valore delle prove di ricerca come uno dei tanti fattori considerati dai responsabili delle politiche e dai professionisti. Ancor di più, nel caso della ricerca medica, l'analisi della produzione scientifica e del suo impatto è indispensabile, considerando le sue implicazioni per la*

[1] Massimo Aria, Department of Economics and Statistics, University of Naples Federico II, Naples, Italy; email: massimo.aria@unina.it

[2] Corrado Cuccurullo, Department of Economics, University of Campania Luigi Vanvitelli, Caserta, Italy; email: corrado.cuccurullo@unicampania.it

[3] Luca D'Aniello, Department of Social Sciences, University of Naples Federico II, Naples, Italy; email: luca.daniello@unina.it

[4] Maria Spano, Department of Economics and Statistics, University of Naples Federico II, Naples, Italy; email: maria.spano@unina.it

salute pubblica. Attraverso l'uso di metodi di science mapping, proponiamo uno strumento per la visualizzazione del posizionamento strategico di diversi Centri medici accademici italiani di proprietà pubblica in termini di posizionamento nella ricerca. La nostra proposta fornisce sia un quadro concettuale per i policy maker, sia uno strumento per orientare le attività di ricerca e le future collaborazioni degli istituti sanitari verso scenari sempre più innovativi.

Key words: bibliometrics, science mapping, academic medical centers, research positioning

1 Introduction

Health research is to be considered one of the key elements and an integral part of the complex of activities carried out by the National Health Service (SSN). In Italy, public institutions dealing with health research are known as Academic Medical Centers (AMCs - that is 20 public AMCs as "Aziende Ospedaliere integrate con l'Università" (AOUs), 9 public AMCs as "Ex Policlinici Universitari a gestione diretta" (AOUs SSN), 21 public-owned "Istituti di Ricovero e Cura a Carattere Scientifico" (IRCCS) (Ministry of Health - http://www.salute.gov.it/, 2018)). The aim of health research is a continuous improvement of assistance, care and services that significantly increases the health of citizens, their expectations of well-being, and quality of life. In addition to increasing scientific knowledge a good research activity has a great impact on cultural and professional growth of researchers. It facilitates them in entering international research networks and contributes to increasing the prestige of the involved healthcare institutions.

Health research must be considered as an investment for the future and then, the analysis of scientific production and its impact becomes indispensable. Bibliometrics introduces transparent and reproducible methods for measuring the quantity and quality of scientific production (*performance analysis*) (Cuccurullo *et al*., 2016). Moreover, it provides a conceptual structure of the extant research by synthesizing past research findings, deducing trends and gaps, and identifying the main centers of interest (*science mapping*) (Zaho, 2010).

In this work, we focus on science mapping as it allows identifying and displaying themes and trends with a synchronic (Callon et al., 1983) or a diachronic perspective (Cobo et al., 2011). By means of science mapping techniques, namely the term co-occurrence networks, and strategic/thematic maps, we aim at providing a data visualization of strategic positioning of different Italian public owned AMCs in terms of their research positioning.

We identify the research-front of different AMCs and then, we visualize them in a joint representation, useful for comparing their main research themes and at the same time their different specializations simultaneously. Moreover, this innovative approach is useful also to detect and capture their research evolution during the years (Cuccurullo et al., 2021).

Mapping the dynamic positioning of Italian medical research at various levels (i.e. national, regional, AMCs type, AMC) will provide a conceptual framework for policymakers and managers to understand and manage the issues of the AMCs (e.g. appropriate funding mechanisms for financing the triple-mission). The tool we propose could be useful for the institutions themselves to direct their research efforts towards increasingly innovative fronts taking into account the general landscape and at the same time exploiting this information to establish collaborations with other AMCs dealing with the same research topics.

We propose a focus on the evolution of scientific production published the last 20 years of AOUs located in metropolitan areas.

This work has been partially financed by the research project "Leading Change in Academic Medical Centers", funded by the competitive call for projects V:ALERE 2019. The project aims to provide evidence, advice, and remarks to support System and AMC decision-makers to address the many challenges that AMC faces.

2 Methodological framework

To map the conceptual structure of each AOU we conducted two related analyses: a term co-occurrence network analysis and a strategic or thematic map. The combined use of these techniques allows us to illustrate: how terms relate to each other, the main research themes within each institution, and how they develop.

The basic idea behind the term co-occurrence network analysis (Wang et al., 2019) is that each research field or topic can be represented as a set of terms (e.g. keywords, terms extracted from titles, or abstracts). Network representation is used to understand the themes covered by a research field, to define which are the most important and the most recent research fronts. Following the network approach, we built a term co-occurrence matrix, in which each cell outside the principal diagonal contains the number of times two terms appear together in the articles (co-occur). Then, the co-occurrences among terms were normalized by the association index as proposed by Van Eck and Waltman (2009). This measure assumes values in the interval [0,1] and reflects the strength of the association among terms. Co-occurrence matrices can be seen as undirected weighted graphs; therefore, we can build a network in which each term is a node and the association between linked terms is expressed as an edge, visualizing both single terms and subsets of terms frequently co-occurring together. To detect subgroups of strongly linked terms, where each subgroup corresponds to a center of interest or to a theme of the analyzed collection, we refer to community detection algorithms (Fortunato, 2010) by using Louvain algorithm (Blondel et al., 2008).

Strategic or Thematic map (Cobo et al., 2011) allows plotting the themes, identified through community detection, in a bi-dimensional matrix where axes are functions of the Callon centrality and density, respectively (Callon et al., 1983). Centrality can be read as the importance of the theme in the research field, while density can be read as a measure of the theme's development.

In this way, we identified the conceptual structure of each AOU in the three different temporal intervals. Then, we standardized centrality and density values, in order to make a comparison among the research fronts of the different institutions by plotting themes in a joint map. The obtained strategic map allows defining four typologies of themes (Cahlik, 2000) according to the quadrant in which they are placed. Themes in the upper-right quadrant are known as the motor themes. They are characterized by both high centrality and density. This means that they are both developed and important for the research field. Themes in the upper-left quadrant are known as isolated themes or niche themes. They have well developed internal links (high density) but unimportant external links and so are of only limited importance for the field (low centrality). Themes in the lower-left quadrant are known as emerging or declining themes. They have both low centrality and density meaning that are weakly developed or marginal. Themes in the lower-right quadrant are known as basic and transversal themes. They are characterized by high centrality and low density. These themes are important for a research field and concern general topics transversal to the different research areas of the field. In each temporal interval, we build the strategic maps using the KeyWords Plus (ID) as units of analysis. The ID are words or phrases that frequently appear in the titles of an article's references but do not appear in the title of the publication itself. Their generation is based upon a special algorithm (Garfield, 1990) that is unique to Clarivate Analytics databases.

3 Data collection and Main findings

As we said above, in Italy there are 20 public AOUs. To make a comparison among homogeneous institutions, in this work, we considered only the six institutions located in metropolitan areas. To retrieve their scientific production the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was used (Liberati et al., 2009). We queried the Web of Science (WoS) indexing database – launched by the Institute for Scientific Information (ISI) and now maintained by Clarivate Analytics – all the publications from January 2000 to December 2019. To identify the publications related to each AOU, we searched by full name, part of the organization name's or by its commonly known abbreviation from the Organizations – Enhanced List available on WoS (e.g. "Azienda Ospedaliera Universitaria (AOU) MEYER" for the "Azienda Ospedaliero-Universitaria Meyer"). We limit our search by document type and selected only Articles, Proceedings Papers, Review Articles, and Book Chapters in the English language. The records were exported into PlainText format.

Starting from our final collection, we loaded the data and converted it into R data frame using bibliometrix, an open-source tool for quantitative research in scientometrics and bibliometrics that includes all the main methods for performance analysis and science mapping (Aria and Cuccurullo, 2017).

To highlight the main research themes of AOUs and evaluating their evolution over time, we decided to divide our reference timespan (2000–2019) into three-time slices. In all AOUs there was a constant increase of the scientific production, both in the total number and in the average number of publications per author.

In Figures 1-2-3 the thematic evolution of AOUs' research is shown. It is worth noting that each theme, identified with the community detection algorithm, is labelled with the corresponding most frequent ID. In the first time slice (2000 – 2006) for almost all AOUs the theme *expression* is a motor theme, highlighting how their research focuses on investigating genetics of diseases. This theme disappears in the second time slice (2007 – 2013) for some AUOs and for all of them in the third slice (2014 – 2019). Since 2000 studies focus on *management* that appeared as an emerging theme only for the AOU CAREGGI on the lower-left quadrant - low density and low centrality. In the second period, *management* becomes a motor theme for many AOUs, and then starting to shift from the upper-right quadrant to the lower-right quadrant in the third slice (2014 – 2019), consolidating its role as traditional theme - low density and high centrality. This is an interesting result as it shows how AOUs research is evolving considering how important the management and organizational aspects of the institutions is to guarantee efficiency in care services. On the upper-left quadrant, we have observed that niche themes - low centrality and high density - have increased over time. This means that the research of AOUs is oriented towards studies more and more specialized from 2000 to 2019.

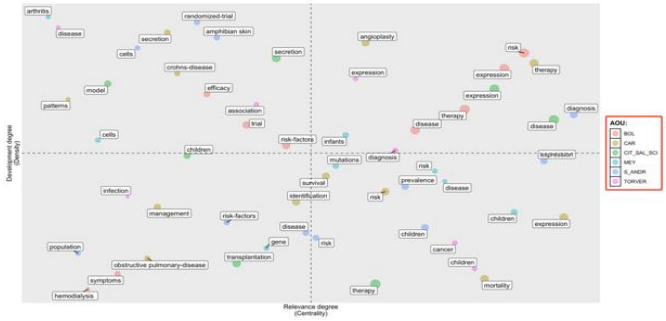


Figure 1: Thematic map **slice 1**:2000-2006

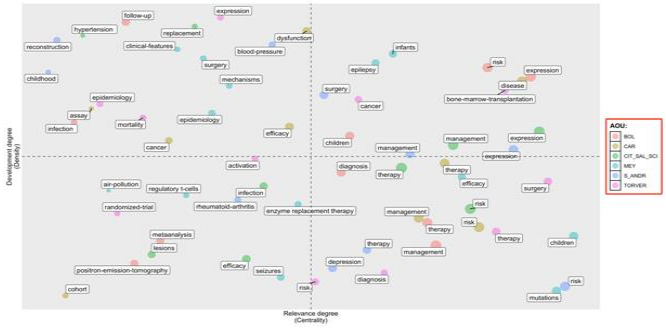


Figure 2: Thematic map **slice 2**:2007-2013

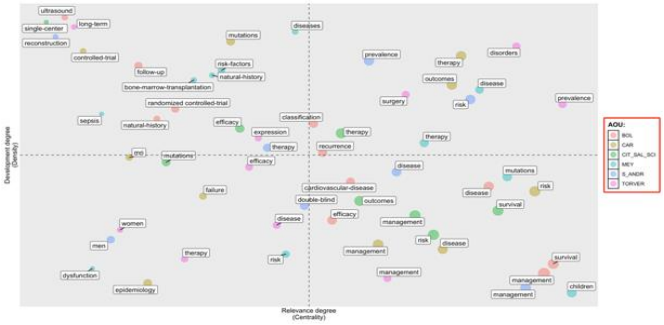


Figure 3: Thematic map **slice 3**:2014-2019

These graphical representations summarize many aspects of AOUs research. Obviously, the presented results are only a small part of what could be observed starting from the thematic maps. Therefore, they are powerful decision support tools for the different agents involved in the health system.

References

1. Aria, M., Cuccurullo, C.: bibliometrix: An R-tool for comprehensive science mapping analysis. J. Informetr. **11** (4), pp. 959-975, (2017)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech-Theory E., 2008. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008, (2008)
3. Cahlik, T.: Comparison of the maps of science. Scientometrics **49** (3), pp. 373-387, (2000)
4. Callon, M., Courtial, J.P., Turner, W. A., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis. Soc. Sc. Infor. **22** (2), pp. 191-235, (1983)
5. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. J. Informetr. **5** (1), pp. 146–166, (2011)
6. Cuccurullo, C., Aria, M., Sarto, F.: Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains. Scientometrics **108** (2), pp. 595-611, (2016)
7. Cuccurullo, C., D'Aniello, L., Spano, M.: Thematic atlas of Italian oncological research: the analysis of public IRCCS. In ASA 2021 Statistics and Information Systems for Policy Evaluation: Book of short papers of the opening conference **127,** pp. 97-103. Firenze University Press (2021)
8. Eck, N.J.V., Waltman, L.: How to normalize cooccurrence data? An analysis of some well-known similarity measures. J. Am. Soc. Inf. Sci. Tec. **60** (8), pp. 1635-1651, (2009)
9. Fortunato, S.: Community detection in graphs. Phys. Rep. **486** (3), pp. 75-174, (2010)
10. Garfield, E.: Keywords Plus®: ISI's breakthrough retrieval method. Part 1. Expanding your searching power on Current Contents on Diskette, Curr. Contents **32**, pp. 5-9, (1990)
11. Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. **62** (10), pp. e1-e34, (2009)
12. Wang, H., Zhao, Y., Dang, B., Han, P., Shi, X.: Network centrality and innovation performance: the role of formal and informal institutions in emerging economies, J. Bus. Ind. Mark. **34** (6), pp. 1388-1400, (2019)
    Zhao, D.: Characteristics and impact of grant-funded research: a case study of the library and information science field. Scientometrics **84** (2), pp. 293-306. DOI= 10.1007/s11192-010-0191-y, (2010)

Session of free contributes SCL3 – *Society and Tourism*
Chair: Emma Zavarrone

Local clustering coefficient to measure intra-regional tourism in Italy

Coefficiente di clustering locale per misurare i flussi turistici tra le regioni italiane

Roberto Rondinelli, Lucio Palazzo and Riccardo Ievoli

Abstract Network analysis represents a useful approach to measure intra-regional tourism flows. Relevant information about this phenomenon can be retrieved using measures of local clustering. We analyse Italian tourism flows in two-year period 2018-2019, identifying different patterns and two main groups of regions.
Abstract *L'analisi delle reti rappresenta un utile approccio per misurare i flussi turistici tra regioni. In questo quadro, misure di clustering a livello locale possono essere utilizzate per ottenere informazioni rilevanti. Di seguito si analizzano i flussi turistici in Italia nel periodo 2018-2019 identificando diversi profili regionali, classificabili in due gruppi principali.*

Key words: tourism flows, network analysis, tourism circle

1 Introduction

The analysis of intra-regional tourism flows represents a wide area of study, especially from a statistical perspective. Recent approaches include spatio-temporal modelling [1] with an inferential aim, or in-depth descriptive analysis through clustering of georeferenced data [6]. In Italy, the tourism is one of the greatest industry and, in this regard, the competitiveness, attractiveness and receptivity of Italian regions can be influenced by their prosperity. Moreover, also the cultural heritage [3] plays a relevant role, especially regarding leisure and holiday trips. For these rea-

Roberto Rondinelli
University of Naples Federico II, roberto.rondinelli@unina.it

Lucio Palazzo
University of Naples Federico II, lucio.palazzo@unina.it

Riccardo Ievoli
University of Ferrara, riccardo.ievoli@unife.it

sons, aggregate studies concerning trips and features of the travellers can be placed by a more structural analysis of the regional flows.

Network Analysis contains a broad range of methods that can help to unveil unobservable patterns and to produce summary network-based measures presenting a meaningful interpretation for policy makers and experts. Trips can be expressed through a network relating the territories of a country. In particular, the analysis of network topology for the tourism flows can help arising not only the receptivity (and/or attractiveness) of a single region, but also its "position" in the connectivity among the others. Descriptive network analysis has been recently applied to the tourism flows in case of large destination systems [5]. Although groups of regions can be identified through spatial proximity or other conventional characteristics (such as environmental, cultural or political features), we focus on the topological relationships based on the interconnectivity of triplets of regions (triadic analysis). Starting from these premises, we identify the following research questions:

RQ.1: *What are the network-based regional characteristics in terms of receptivity/attractiveness?*

RQ.2: *What is the level of network clustering of each region, according to different directionality flows?*

To answer these questions, the aim of this paper is to explore the potential of the Local Clustering Coefficient (LCC) in case of directed and weighted network of the tourism flows. This approach is carried out considering the intra-regional tourism flows of Italy. The data for the application come from the "trips and holidays" survey of the Italian National Institute of Statistics (ISTAT). Descriptive usage of network-based clustering coefficients in tourism has been applied in recent works [2, 8].

2 Data and Methodology

The trips between Italian regions in the two-year period 2018-2019[1] are considered: data come from the survey "trips and holidays"[2], included in the wider "households budget" survey of ISTAT. We consider the intra-regional trips for holiday reasons, aggregating the flows through the following information: *a*) the region of residence and *b*) the destination region of the households. Holiday trips are defined as the travels for the following motivations: *i*) leisure *ii*) visits to relatives and friends *iii*) religion *iv*) well-being treatments (including spa). We normalize the flows using the survey's weights to take into account differences in regional populations.

The applied methodology is related to directed and weighted networks. Here, a network G is defined as an ordered triple (V,E,W), where V is a set of vertices (or nodes) v, $E \subseteq V \times V$ is a set of edges (or links) e, and W is a set of weights ω

[1] The following year is not considered because data are highly influenced by the COVID-19 widespread

[2] https://www.istat.it/en/archivio/227020

associated to the edges. The mapping $\omega : E \to \mathbb{R}$ defines the weight related to each edge. A network can be expressed in the form of an *adjacency matrix* $A = (a_{ij})$ (not necessarily symmetrical), with $a_{ij} = \omega(e_{ij})$ if $\exists\, e_{ij} \in E$, and $a_{ij} = 0$ otherwise.

For each considered year, network vertices are the $n = 20$ Italian regions. To normalize for the different travel volumes, we decided to study the relative travel flows. Given a departure region v_i, i.e. the region of residence of households, and an arrival region v_j, a directed weighted edge p_{ij} expresses the percentage of outgoing trips to v_j over the total trips belonging to the ingoing region v_i:

$$p_{ij} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}. \tag{1}$$

The resulting data matrix P (with elements p_{ij}) allows to study the tourism flows in terms of normalized ratios of outgoing trips, emphasizing the choice of the favourite destination region.

Given a region v_i, there are two possible flows: outgoing and ingoing. The outgoing flow is determined by the trips towards the other regions v_j (with $j \neq i$), while the ingoing ones are given by the arrivals from regions v_j (with $j \neq i$) to region v_i. The final set of local (outgoing and ingoing) connections of each region v_i, can be defined as its *tourism circle* (TC_i).

Besides the number of outgoing and ingoing relationships (and their volumes) associated to each region v_i, a TC_i includes the relationships between the network neighbours of v_i, thus any triangular tourism flow in which v_i can be involved. To this end, the level of clusterization of v_i in terms of triadic tourism flows is consistently measured by the LCC_i. In the case of directed and weighted networks, LCC has been defined by [4] by the following expression:

$$LCC_i^D(P) = \frac{[P^{[1/3]} + (P^T)^{[1/3]}]_{ii}^3}{2[d_i^{tot}(d_i^{tot} - 1) - 2d_i^{\leftrightarrow}]}, \tag{2}$$

where $LCC_i^D(P)$ stands for directed (D) and weighted (P) LCC of vertex v_i, P is the weighted matrix previously normalized[3], d_i^{tot} is the total degree $d_i^{tot} = d_i^{in} + d_i^{out}$, with $d_i^{in} = \sum_{j \neq i} e_{ji}$ and $d_i^{out} = \sum_{j \neq i} e_{ij}$, and $d_i^{\leftrightarrow} = \sum_{j \neq i} e_{ij} e_{ji}$ are the bilateral edges in which the vertex v_i is involved. This general formulation is specified for different directed triadic configurations: *In*, *Out*, *Cycle*, *Middleman* (refer to Table 1 in [4]).

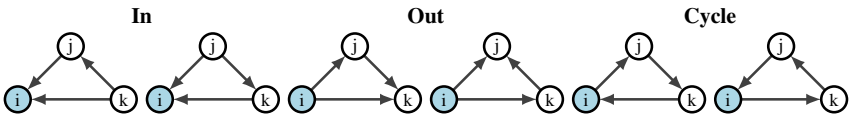


Fig. 1 Triadic configurations

[3] In this case, weighted edges are normalized dividing by the maximum value of them. Since we already normalized, here it is not necessary.

Basically, the *In* configurations define the level of clustering of region v_i within its TC_i according to the ingoing tourism flows towards it, while the *Out* configurations has the same interpretation but considering the outgoing flows. Finally, the *Cycle* configurations consider the cyclic interchange of trips. Given the high correlation between the observed LCC of the *Cycle* and the *Middleman* configurations, we decided to use only the former due to its meaningfulness in the tourism flows. Figure 1 shows the aforementioned triadic configurations.

3 Results and Discussion

Firstly, we present the descriptive statistics of the network regarding Italian tourism flows, to answer to the first research question *RQ.1*. In Table 1, d^{in} and d^{out} are the number of inbound and outbound regions linked to each region v_i, Mode^{out} is the favourite destination for the outbound trips of each region v_i, while $loop^{in}$ and $loop^{out}$ are the proportions of inbound and outbound trips carried out in the same region of residence, respectively. Despite the high connectivity between Italian regions, we notice that are some cases (Valle D'Aosta, Molise and Basilicata) that generally are chosen as trip destination by the inhabitants of few regions.

Table 1 Descriptive analysis of network-related measures for Italian regions

Region	d^{in}		d^{out}		$loop^{in}$		$loop^{out}$		Mode^{out}	
	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
Piemonte	17	15	17	15	0.30	0.39	0.14	0.22	Liguria (0.17)	Liguria (0.27)
Valle D'Aosta	6	7	12	14	0.00	0.01	0.00	0.05	Liguria (0.19)	Liguria (0.22)
Lombardia	18	19	18	17	0.34	0.27	0.15	0.12	Emilia Romagna (0.16)	Liguria (0.16)
Trentino Alto Adige	14	11	17	17	0.03	0.06	0.13	0.20	Veneto (0.31)	Trentino Alto Adige (0.20)
Veneto	19	16	16	19	0.30	0.37	0.25	0.28	Veneto (0.25)	Veneto (0.28)
Friuli Venezia Giulia	11	9	15	12	0.27	0.33	0.29	0.35	Friuli Venezia Giulia (0.29)	Friuli Venezia Giulia (0.35)
Liguria	15	12	14	18	0.11	0.11	0.18	0.17	Liguria (0.18)	Toscana (0.19)
Emilia Romagna	18	18	17	18	0.30	0.25	0.24	0.18	Emilia Romagna (0.24)	Emilia Romagna (0.18)
Toscana	19	18	18	18	0.36	0.27	0.42	0.35	Toscana (0.42)	Toscana (0.35)
Umbria	14	11	14	15	0.01	0.03	0.01	0.02	Lazio (0.21)	Emilia Romagna (0.24)
Marche	14	14	14	13	0.08	0.04	0.08	0.07	Trentino Alto Adige (0.14)	Veneto (0.16)
Lazio	19	19	18	17	0.21	0.30	0.15	0.22	Toscana (0.19)	Lazio (0.22)
Abruzzo	13	13	14	8	0.04	0.10	0.08	0.25	Lazio (0.18)	Abruzzo (0.25)
Molise	4	6	16	12	0.04	0.07	0.01	0.10	Puglia (0.21)	Abruzzo (0.20)
Campania	19	16	15	14	0.28	0.27	0.25	0.20	Campania (0.25)	Campania (0.20)
Puglia	16	17	11	13	0.06	0.14	0.23	0.40	Puglia (0.23)	Puglia (0.40)
Basilicata	7	10	9	9	0.20	0.06	0.15	0.17	Puglia (0.26)	Piemonte (0.21)
Calabria	15	13	11	7	0.11	0.03	0.28	0.09	Calabria (0.28)	Lazio (0.27)
Sicilia	16	16	12	11	0.32	0.30	0.37	0.36	Sicilia (0.37)	Sicilia (0.36)
Sardegna	13	14	9	7	0.23	0.34	0.36	0.59	Sardegna (0.36)	Sardegna (0.59)

On the other hand, residents of Sardegna, Calabria and Basilicata are more likely to choose fewer trip destinations with respect to residents of other regions, even if some differences between consecutive years arises (see e.g. Calabria). Concerning Mode^{out}, about half of the regions are the favourite destinations of the their own residents. Combining the results of Mode^{out} and $loop^{in}$, many exceptions arise, e.g. the favourite destination of the Piemonte's residents is the closest "seaside" region of Liguria, while the inhabitants of Molise prefer to travel towards bordering regions such as Puglia and Abruzzo.

Local clustering coefficient to measure intra-regional tourism in Italy

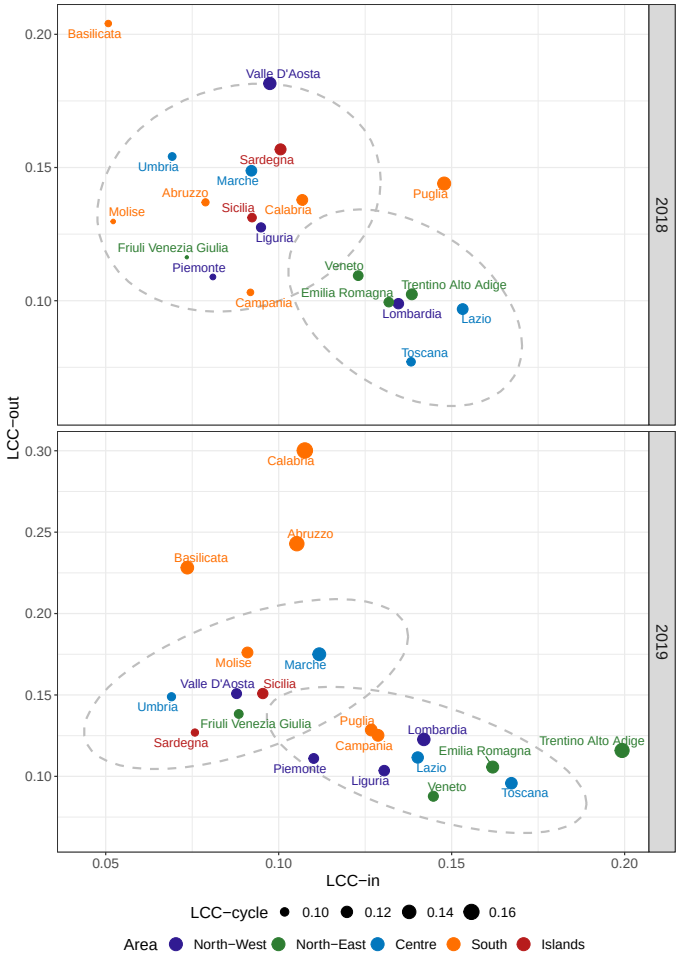


Fig. 2 Distribution of Italian regions according to their LCC_{in}, LCC_{out}, LCC_{cycle} (the size of the points is related to the observed LCC_{cycle}) and geographical area for 2018 (upper panel) and 2019 (lower panel) tourism flows.

For what concerns the second research question (*RQ.2*), Figure 2 helps to map the regional differences in terms of LCC with respect to their geographical areas (North-West, North-East, Centre, South, Islands) in the two consecutive years (2018-2019). Observing the distribution of the outgoing LCC, it can be noticed that its range increases in the 2019 while the range of ingoing LCC has a slight decrease.

Two main groups, emphasized by the dashed ellipses, can be observed:

Weak inbound and medium outbound TC: this group includes regions with low level of LCC^{in} and medium level of LCC^{out}. The low level of ingoing clustering in terms of TC_i can be due either to a sort of exclusivity of the trips or weak

ingoing triadic flows. Regarding the outbound trips, the regions in a TC_i appear to be connected each other with stronger tourism flows.

Strong inbound and weak outbound TC: this group includes regions with high level of LCC^{in} and, at the same time, low level of LCC^{out}. The interpretation of the flows inside of the regional TCs can be opposed to the previous one.

In addition, some peculiar cases are located in the top-left side of both panels: Basilicata in 2018 and Abruzzo, Calabria and Basilicata in 2019. Their low level of LCC^{in} is in line with the first group, while the high level of LCC^{out} results in a strong connectivity within their TCs. In general, LCC^{in} is lower for Southern regions and Islands, denoting weaker TCs' connectivity or an increased characterization of their ingoing patterns. In the 2019 panel, Trentino Alto Adige accounts for a special case: its ingoing TC has a really strong level of intra-regional tourism flow.

Even the LCC^{cycle} provides useful information. High scores of LCC^{cycle} are related with hybrid ingoing and outgoing composition of the TC_i. Some regions show propensity of their residents to travel out of their region, but also a good inbound attractiveness. Cyclicity of the tourism flows arises for the North-Eastern regions and Lombardia, and for some of the Southern Regions (Puglia and Calabria) in both years, whereas it holds only for one year in Campania, Molise and Piemonte.

To conclude, the empirical analysis shows the potential of these network-based local clustering measures to highlight patterns of intra-regional tourism flows, identifying two main regional profiles of the tourism circle. From an evaluation perspective, the proposed approach can be used to study the effects of tourism-related policies on the attractiveness/receptivity of a region or a group of regions. Presented results should be the subject of further studies, especially because these patterns are related to geographical, climatic, cultural and political aspects here not discussed.

References

1. Álvarez-Díaz, M., D'Hombres, B., Ghisetti, C.: Modelling inter-and intra-regional tourism flows in Spain–a spatial econometric approach. Reg. Stat., **7**(2), 3-34 (2017).
2. Baggio, R.: Tourism destinations: A universality conjecture based on network science. Ann. Tour. Res., **82**, 102929 (2020).
3. De Simone, E., Canale, R. R., Di Maio, A.: Do UNESCO World Heritage Sites influence international tourist arrivals? Evidence from Italian provincial data. Soc. Indic. Res., **146**(1), 345-359 (2019).
4. Fagiolo, G.: Clustering in complex directed networks. Phys. Rev. E, **76**(2), 026107 (2007).
5. Kádár, B., Gede, M.: Tourism flows in large-scale destination systems. Ann. Tour. Res., **87**, 103113 (2021).
6. Majewska, J., Truskolaski, S.: Cluster-mapping procedure for tourism regions based on geostatistics and fuzzy clustering: example of Polish districts. Curr. Issues Tour., **22**(19), 2365-2385 (2019).
7. Watts, D. J., Strogatz, S. H.: Collective dynamics of 'small-world'networks. Nature, **393**(6684), 440-442 (1998).
8. Zhu, H. (2021): Multilevel understanding dynamic changes in inbound tourist flow network (ITFN) structure: topology, collaboration, and competitiveness. Curr. Issues Tour., **24**(14), 2059-2077 (2021).

Challenges And New Jobs for the Post-Covid Tourism: the Perspective of Tourism Managers

Sfide nel post Covid: il punto di vista degli operatori del tursimo

Emma Zavarrone, Martha Friel and Alessia Forciniti

Abstract The Covid-19 pandemic has had profound impacts on the tourism industry, giving businesses and destinations new challenges and priorities to sustain and promote tourism competitiveness. Tourist human capital plays a fundamental role in defining these challenges. By means of a logistic regression model, the paper explores the human capital demo-social determinants that impact on the definition of tourism challenge factors in the pre and post Covid-19. A number of social and demographic determinants of resilience for the tourism sector are also analysed through a Latent Class Analysis suggesting some policy implications directions for future research in the area.

Abstract *La pandemia di Covid-19 ha avuto profondi impatti sull'industria del turismo, dando alle imprese e alle destinazioni nuove sfide e priorità per sostenere e promuovere la competitività del turismo. Il capitale umano turistico svolge un ruolo fondamentale nella definizione di queste sfide. Attraverso un modello di regressione logistica, il paper esplora le determinanti demo-sociali del capitale umano che impattano sulla definizione dei fattori di sfida turistica nel pre e post Covid-19. Anche un numero di determinanti socio-demografiche della resilienza per il settore del turismo sono state analizzate attraverso la Latent Class Analysis, suggerendo alcune indicazioni sulle implicazioni politiche per la ricerca futura in questo settore.*

Key words: tourism human capital, logistic regression, latent class analysis, resilience

[1] Emma Zavarrone, IULM University; emma.zavarrone@iulm.it
Martha Friel, IULM University; martha.friel@iulm.it
Alessia Forciniti, University of Naples Federico II; alessia.forciniti@unina.it

1 Introduction and background

The central role that human capital plays in the development of competitiveness of the global tourism industry is widely recognized in literature and practice.

An educated and skilled labour force is a critical dimension for the successful design and delivery of competitive tourism offerings. Despite this centrality of the issue of people and of human capital in tourism, the tourism industry, compared to other sectors, has manifested over the last decades a series of criticalities linked to the attractiveness and qualitative development of its labour market.

In particular, the tourism industry presents a series of structural problems which, in turn, depend on the organization of the tourism business system characterized by the very high presence of SMEs and therefore by the reduced capacity and resources to facilitate and encourage workforce development.

Among these structural problems, academic literature as well as reports by national and international agencies (Baum, 2013; OECD, 2014; Stacey *et al*., 2015), informed the high dimension of seasonal work, the wage gap compared to other sectors, recruitment and retention difficulties, high turnover and vacancy rates, poor image and weak training culture.

The advent of the Covid-19 pandemic has made these reflections even more urgent, on the one hand due to the devastating impact of the pandemic on the tourism industry and on its workforce, on the other, due to the need for an overall redesign of the offer at the destination and company level. Consequently, it has become a priority to know and measure the changes in values, behaviours, needs, skills and expectations of people working in tourism. This to generate the conditions for making the most of it from the tourist human capital and to make it a central interlocutor on the global challenges and priorities of the sector for the post covid-19.

Tourism literature has extensively investigated the issue of human capital by adopting a multiplicity of perspectives that also depends on the wide variety of economic activities that make up the tourism industry (Rey-Maquieira, Tugores and Ramos, 2007).

However, a lack of literature that connects the characteristics of tourism human capital in terms of education and work experience with the vision for the future development of the sector still persists. This vision becomes even more interesting to analyse after covid-19, in consideration (i) of the great impact that the pandemic has had on the tourist workforce and on working conditions in tourism, (ii) of the challenges posed to businesses and destinations in terms of competitiveness recovery.

The paper explores two main research questions:

(RQ1) What are the pre- and post-covid challenge factors and expectations identified by tourism managers?

(RQ2) How can this expectation be clustered with respect to the characteristics of the human capital of tourism?

2 Data

To investigate the two above mentioned research questions, we conducted secondary data analysis using data by Giaccardi & Associati – an Italian consultancy company specialized in tourism marketing – that were collected before and after the Coronavirus outbreak for investigating the different dimensions related to the success in the tourism sector.

Data were gathered from hotels' owners and managers of regional tourism authority that are customers of the company. A stratified random sample procedure was used to ensure a random selection of participation for each wave, the first sample (2019) was composed by 336 units and the second one (2020) by 290. The questionnaires between the waves have kept several items enough if, due to the Coronavirus, new questions were added. Each questionnaire has composed by 24 items divided in five sections: challenge factors, expectations, motivations, future and perspective in tourism beyond demo-social aspects. We focused on the challenge factors (CF) and the social demographic variables for each questionnaire. The CF can be considered such as a proxy of resilience, a latent variable. A challenge factors question has been proposed with multiple choices related to the different aspects of tourism dynamics. We have transformed each choice in a binary variable. We selected only the three binary variables characterized by a high number of respondents. The selected CF variables for 2019 were: innovation in the tourism offer and experience, destination's brand reputation and competition quality training. The selected 2020 CF variables were: innovation in the tourism offer and experience, destination's brand and reputation and security standards.

3 Methodology

The RQ1 has been solved comparing logistic regression models (Tab.1).

The 2020 logistic models select innovation (V1) and health (V2) as two "candidate" proxies of the resilience construct. To satisfy RQ2, we compare three Latent Class models for identifying profiles of social and demographic determinants of resilience. Latent Class Analysis (LCA) introduced by Lazarsfeld (1950) can be considered as a subset of structural equation modelling allowed to identify sub-groups of operators called latent classes that share the same characteristics (Hagenaars & Mccutcheon, 2002) on the basis of categorical models of responses to observed variables (Muthén & Muthén, 2000; Wolke et al., 2013). LCA main advantages is to assess model fit and determine an appropriate number of latent classes by means of goodness of fit statistics. Three LCA models can be set up as latent regression logit model with J latent classes with binary dependent variable Y can be write as (Holm and Pedersen, 2007):

$$P(Y_i = (1|X)) = \sum_{j=1}^{J} P(Y_i = (1|X), \Xi = \epsilon_j) P(\Xi = \epsilon_j)$$
$$= \sum_{j=1}^{J} \frac{exp(\alpha + \beta X + \epsilon_j) P(\Xi = \epsilon_j)}{1 + exp(\alpha + \beta X + \epsilon_j)}$$

i= 1 =Innovation, 2 = Health, 3= V4
j = 1,..,J, where J is the number of latent classes

where

A	constant term
X	matrix of explanatory variables
β	matrix of regression coefficients
ϵ_j	effect of the j'th latent class on the probability of observing Y = 1
P($\Xi = \epsilon_j$)	Frequency of the *j*'th latent class in the population

V4 is a new variable composed by the conjoint recode of the resilience proxies with these values:

$$V4 = \begin{cases} 0 \ \ if\ Innovation\ and\ Health = 0 \\ 1 \ \ if\ Innovation\ = 1\ and\ Health = 0 \\ 2 \ \ if\ Innovation = 0\ and\ Health = 1 \\ 3 \ \ if\ Innovation = 1\ and\ Health = 1 \end{cases}$$

4 Results and discussion

Descriptive statistics for 2019 and 2020 have highlighted in both years there is a prevalence of male respondents (53%).

The level of specialization in the tourism sector increased from 23 to 37% from 2019 to 2020. The experience greater than of 20 years has grown from 35 to 41%, while the role as owner has decreased by 19%. The future view passed from positive for 82% of respondents to undefined for 57% of them. The result of logistic regression models on CF allowed to observe the transition from the market strategies based on factors such as reputation and quality of skills in 2019 to factors such as innovation and health in 2020 (Table 1). The result of the logistic regression on CF allowed to observe a shift of focus between 2019 and 2020: from a more market-oriented one in 2019, based on factors such as destination's branding and quality of competitors, to one more centred on innovation and health in 2020. In response to RQ2, the model for classes that best explains this dimension is that of innovation, in according to the goodness of fit statistics AIC, BIC and $\chi 2$, which showed lower values than the other models analysed. The same statistics allowed to determine 2 latent classes as optimal

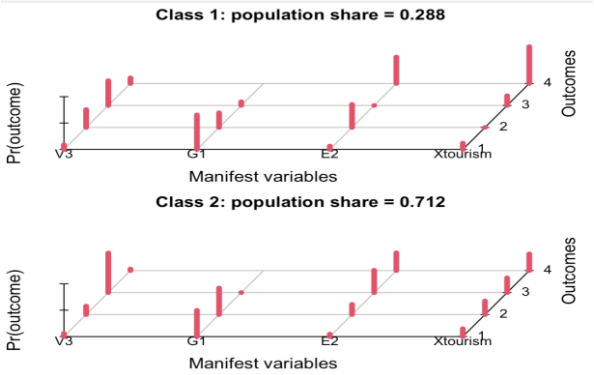
number of groups which range from 29 to 71% of the sample, as shown by the graphic representation (Figure 1) in which the red bars identify the conditioned probability of the respondent, for latent class, to belong to a category. The first, and smallest, class (29%) showed main interest towards the tourism experience. The second, and largest, class (71%) instead displayed greater attention for the higher education and innovation to generate resilience. Therefore, a dichotomy has emerged in terms of CF and resilience: a class is dominated by the specialization and see the higher education as an engine of resilience, another class is conservative and dominated by experience.

Table 1: Logistic models: comparisons

Variables	**Challenge factors**			
	Innovation 2020	Health 2020	Reputation 2019	Quality of compet 2019
Age	-0.01	0.03*	0.01	-0.01
Gender	-0.63**	0.30	0.47*	-0.27
Education-Bachelor	-0.13	0.03	-0.55*	0.26
Education-Degree	-0.01	-0.46	-0.66	0.08
Experience tourism <5 years	1.02*	-0.68		0.18
5 years<Experience tourism <10 years	1.35**	-1.98***		0.51
5 years<Experience tourism <10 years	1.38**	-1.34*		0.819441**
Role-Owner	-0.65*	0.64		
Role-technicians	0.24	-1.26*		0.750843*
Role Public	0.99**	-0.29		
Future View-Very Negative		0.32	1.55	1.75
Future View-Negative		-1.18*	1.55	1.18
Future View-Positive		-0.29	1.46	1.28
Region -North Italy		-0.12	0.59**	-0.70181*
Region Center Italy		-0.03		
Pseudo R2	0.73	0.61	0.51	0.49

* = $p \leq 0.10$; * * = $p \leq 0.05$; *** = $p \leq 0.01$

Figure 1: Latent Class Analysis



References

1. Baum, T.: International perspectives on women and work in hotels, catering and tourism. Working Paper 1 / 2013. Geneva: ILO (2013)
2. Baum, T.: Human resources in tourism: Still waiting for change? - A 2015 reprise. Tourism Management **50**, 204--212 (2015).
3. Berlin, K. S., Williams, N. A., and Parra, G. R.: An introduction to latent variable mixture modeling (part 1): overview and cross-sectional latent class and latent profile analyses. J. Pediatr. Psychol. **39**, 174--187 (2013).
4. Elsharnouby, T. H., & Elbanna, S.: Change or perish: Examining the role of human capital and dynamic marketing capabilities in the hospitality sector. Tourism Management **82**, (104184) (2021)
5. Hagenaars, J. A., McCutcheon, A. L.: Applied latent class analysis. Cambridge University Press (2002)
6. Holm, A., & Pedersen, M.: Latent class binary regression models–identification and estimation (2007)
7. Kimbu, A. N., Ngoasong, M. Z., Adeola, O., & Afenyo-Agbe, E.: Collaborative networks for sustainable human capital management in women's tourism entrepreneurship: The role of tourism policy. Tourism Planning & Development **16(**2), 161--178 (2019)
8. Lazarsfeld, P. F.: The logical and mathematical foundation of latent structure analysis. In: Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., Clausen, J. A. (Eds.), Studies in social psychology World War II: Measurement and prediction **4**, 361--412. Princeton University Press (1950)
9. Muthén, B. O., Muthén, L. K.: Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. Alcoholism: Clinical & Experimental Research **24**(6), 882--891 (2000)
10. Wolke, D., Copeland, W. E., Angold, A., Costello, E. J.: Impact of bullying in childhood on adult health, wealth, crime, and social outcomes. Psychological Sciences **24**(10), 1958--1970 (2013)
11. Rey-Maquieira, J., Tugores, M., & Ramos, V.: 17 Implications of human capital analysis in tourism. International handbook on the economics of tourism 379 (2007)
12. Stacey J.: Supporting quality jobs in tourism. OECD Tourism Papers, 2015/**02**. Paris, France: OECD Publishing (2015)
13. Thrane, C.: Earnings differentiation in the tourism industry: Gender, human capital and socio-demographic effects. Tourism Management **29**(3), 514--524 (2008).

Sustainable tourism: The case of Albania

Turismo sostenibile: Il caso Albania

Najada Firza

Abstract Thanks to the constant growth of tourism over the years, the tourism sector in Albania has shown a great potential for attractiveness from 2013 onwards. A territory made up of mountain ranges and surrounded by stretches of beautiful coasts, its richness consists in the varied biodiversity of flora and fauna. The challenge for the near future is to maintain the results obtained so far and to increase the competitiveness of tourism in the long term. These objectives must pass through the construction of sustainable tourism which is the key to impress and affirm its presence in the European and world tourism scenarios. Sustainable tourism must be supported by constant monitoring of the local and territorial reality and the construction of sustainability indicators as the directives of sustainable tourism for Europe advocate. The work deals with the construction of some specific sustainability indicators of the sector under the aspect of statistical methodology. To measure the level of sustainability, we start from four main indicators, namely destination management, economic value, social impact and environmental impact. These indicators will be adapted to the Albanian territory.

Abstract *Grazie alla costante crescita del turismo negli anni, il comparto turistico in Albania ha evidenziato un grande potenziale di attrattività a partire dal 2013 in poi. Un territorio fatto di catene montuose e circondato da distese di coste bellissime, la sua ricchezza consiste nella variegata biodiversità di flora e fauna.*

La sfida del prossimo futuro è mantenere i risultati ottenuti fino ad ora ed incrementare la competitività del turismo nel lungo periodo. Tali obiettivi devono passare per la costruzione di un turismo sostenibile che è la chiave per imprimersi ed affermare la propria presenza negli scenari del turismo europeo e mondiale.

[1] Najada Firza, Catholic University "Our Lady of Good Council"; n.firza@unizkm.al

Il turismo sostenibile deve essere supportato da un monitoraggio costante della realtà territoriale e locale e dalla costruzione di indicatori di sostenibilità come auspicano le direttive del turismo sostenibile per l'Europa.
Il lavoro tratta la costruzione di alcuni indicatori specifici di sostenibilità del comparto sotto l'aspetto della metodologia statistica.
Per misurare il livello di sostenibilità si parte da quattro principali indicatori ossia la gestione delle destinazioni, il valore economico, l'impatto sociale e l'impatto ambientale. Tali indicatori saranno adattati al territorio albanese.

Key words: Sustainable Tourism, International tourists, Dynamics of tourism

1 Introduction

In 2013 the European Commission (CE) developed the European Tourism Indicator System (ETIS) for the sustainable management of destinations. This system was created as a voluntary management tool to help destinations to measure and monitor the impact of tourism comprehensively and consequently to make decisions supported by the sustainable management of tourist destinations. A sustainable tourist destination requires the sustainability of the territory it belongs to and policies aimed at encouraging the well-being of the indigenous population in order to make the tourist destination increasingly sustainable. A destination is as welcoming as the well-being of its citizens is high. The work starts from the need to identify, on the basis of statistical data, what has been defined by the European guidelines as a "sustainable tourist destination". What is the starting situation of the tourism sector in Albania? What are the interventions that the government is carrying out to restart tourism and stick to sustainability? To understand this we must carry out an analysis of the territorial context based on the demographic, morphological, economic and structural aspects of Albania. From this it follows the importance of defining specific indicators, able to estimate the level of belonging (to the class of different tourist destinations in the territory) to the "sustainable" tourist destination, among the various municipalities of the territory.

2 Materials and Methods

The indices we have considered are a total of 22. Most of them represent the level that the most significant activities in the area have to participate in the country's GDP. First we have proceeded to normalize the indicators so they have the same range of variation and therefore can assume equal weights and independent of the unit of measurement within the construction of the synthetic index. The analysis

method used is cluster analysis. We initially implemented an exploratory analysis [1] of hierarchical cluster analysis with the Ward method [2]. Subsequently, after removing the unrepresentative indices using ANOVA, a new cluster analysis was used with the non-hierarchical k-means method [3]. Finally we interpreted the results obtained from the analysis.

3 Results, discussion and conclusion

Albania is divided into 12 districts at NUTS 3 level, our analysis divides the territory into 5 groups and to comment on the results it is necessary to mark the belonging of each individual territorial unit to a cluster. In this way we can cross the indicators used for the analysis with the groups obtained in order to discover the characteristics of each group. Figure 1 shows the subdivision of the 12 territorial districts (61 have been grouped into their regions) into groups:

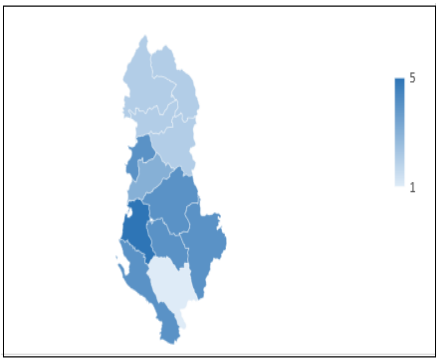


Figure 1: Subdividion of the territory into 5 clusters.

This work aims to be a reflection for the tourism sector and for the territorial tourism policy in order to define personalized strategies based on territorial characteristics. Through the Cluster Analysis we have divided the constituencies into 5 groups. Each group has different characteristics with respect to the activities that contribute most to the national GDP. We summarize these characteristics in the following table.

Table 1: Characteristics for each cluster

Significant Activities	I cluster gastronomic tourism	II cluster coastal/ gastronomic tourism	III cluster cultural tourism	IV cluster potenzial gastronomic/ cultural tourism	V cluster cultural/ gastronomic tourism
Agricolture	high	medium/high	low	high	medium
Arts and e entertainment	medium/low	medium	high	low	medium

				Najada Firza	
Pubblic Sector	medium/high	medium/low	low	low	high
Trade/Trasportation	medium/low	medium	high	medium	high
Buildings	low	medium	high	medium	medium

The first group, formed by the northern districts, deduces that the strategies for the revival of tourism in these districts must focus on the strengths of the predominantly agricultural territory and re-evaluate all those naturalistic landscapes of which this land is proliferating. The profile of the second group, formed by the districts of the south-east and Durres, is made up of cities of coastal tourism and cities of high cultural value. These are areas taken by storm in the summer, but an attractiveness strategy for the whole calendar year should be planned specifically for these landscapes. Furthermore, if we work on improving infrastructures, we could regulate and structure in a sustainable way what is defined as "mass tourism" and furthermore elite tourism could be increased. The third group formed by the district of the capital, Tirana, has different characteristics from the other districts and boasts a significant flow of both foreign and native tourists. Many of these tourist flows are carried out for business reasons. Tirana is a very dynamic city with low rural content and is obviously the hub of the Albanian economy. The strategies that can be used for the capital must leverage the large concentration of cultural heritage existing in Tirana (museums, theaters, opera and other cultural attractions that date back to the period of rigid Albanian communism) but also on a series of natural parks, mountains, beaches and existing farmhouses on the outskirts of the capital. Finally, there are two particular groups also composed of a single constituency: Fier and Gjirokaster. Fier is a district with an important history and three archaeological sites. Mainly agricultural, it has recently become an industrial area with great potential for various commercial activities, especially in the technological sector. Gjirokaster is the smallest Albanian territorial entity and the oldest in the territory. A museum city (proclaimed in 1961 by the Albanian government) which in 2005 became a World Heritage Site. An innate force for cultural tourism that goes well with the culinary one.

References

1. Fabbris, L. (1997) Statistica multivariata. Analisi esplorativa dei dati, McGraw-Hill, Milano.
2. Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58, 236–244.
3. Delvecchio, F. (2010). Statistica per l'analisi di dati multidimensionali, CLEUP, Padova.

Threshold-based Naïve Bayes Classifier: Customer Satisfaction evaluation

Il classificatore Threshold-based Naïve Bayes: valutazione della Customer Satisfaction

Maurizio Romano, Gianpaolo Zammarchi, and Claudio Conversano

Abstract Considering online reviews on Booking.com we propose an ad-hoc classification model for evaluate the Customer Satisfaction, starting from the reviews' content for predicting them as positive/negative. The log-likelihood ratios attributed to each word included in a review are then used to estimate a numeric score. Such a sentiment score, obtained by a review, is then decomposed w.r.t. the different business areas of an hotel for assess the Customer Satisfaction. The proposed approach is evaluated analysing the reviews provided by tourists who stayed in Sardinian hotels.

Abstract *Considerando le recensioni online presenti su Booking.com, si propone un classificatore ad hoc per valutare la Customer Satisfaction che, partendo dal contenuto delle recensioni stesse, le classifica in positive/negative. La log-verosimiglianza attribuita ad ogni parola contenuta in una recensione è successivamente utilizzate per stimare uno score numerico. Tale sentiment-score, ottenuto da una recensione, è successivamente scomposto considerando le differenti aree d'interesse presenti in un hotel al fine di valutare la Customer Satisfaction. L'approccio così proposto è validato analizzando le recensioni fornite da turisti che hanno alloggiato in strutture ricettive sarde.*

Key words: Threshold-based Naïve Bayes Classifier, Customer Satisfaction, Sentiment Analysis, General Sentiment Decomposition, Booking

Maurizio Romano
University of Cagliari, Cagliari, Italy, e-mail: romano.maurizio@unica.it

Gianpaolo Zammarchi
University of Cagliari, Cagliari, Italy, e-mail: gp.zammarchi@unica.it

Claudio Conversano
University of Cagliari, Cagliari, Italy, e-mail: conversa@unica.it

1 Introduction

Textual data have then proved extremely useful, but they are complex, as the language is. For that, many approaches focus more on producing well-performing classifiers and ignore the highly complex interpretability of their models. Instead, we propose a framework able to produce a good sentiment classifier with a particular focus on the model interpretability. Hence, we propose an ad-hoc classification model for evaluate the Customer Satisfaction, starting from the reviews' content for predicting them as positive/negative. Furthermore, we assess an objective sentiment scoring, analyzing the sentiment of people who stayed in Sardinian hotels for evaluating the Customer Satisfaction.

2 The data

For this study, with an ad hoc web scraping Python program, we have collected two separated datasets from Booking.com with a total of 127 features:

- Hotels dataset (86 features): Hotel (3), Review (8), Reviewer (2), Booking's score (11), Accommodation (32), Guest (8), Length of stay (6), Other info (4), Score components (12);
- Comments dataset (41 features): Hotel (2), Comment (6), Reviewer (2), Accommodation (16), Guest (4), Length of stay (3), Other info (2), Score components (6).

More in detail, the web-scraped data, is related to:

- 619 hotels located in Sardinia;
- 66,237 reviews, divided in 106,800 comments (in Italian or English): 44,509 negative + 62,291 positive;
- Period: Jan 3, 2015 – May 27, 2018.

3 Threshold-based Naïve Bayes Classifier

Considering a Natural Language text corpora as a set of reviews $\boldsymbol{r}$ s.t.:

$$r_i = comment_{pos_i} \cup comment_{neg_i}$$

where $comment_{pos}$ ($comment_{neg}$) are set of words (a.k.a. comments) composed by only positive (negative) sentences, and one of them can be equal to $\emptyset$, the basic features of Threshold-based Naïve Bayes classifier applied to reviews' content are as follows. For a specific review r and for each word w ($w \in$ *Bag-of-Words*), we consider the log-odds ratio of w,

$$\begin{aligned} LOR(w) &= \log\left[\frac{P(c_{neg}|w)}{P(c_{pos}|w)}\right] \approx \\ &\approx \log\left[\frac{P(w|c_{neg})}{P(\bar{w}|c_{neg})}\cdot\frac{P(w|c_{pos})}{P(\bar{w}|c_{pos})}\cdot\frac{P(c_{neg})}{P(c_{pos})}\right] = \ldots = \\ &\approx pres_w + abs_w \end{aligned}$$

where $c_{pos}(c_{neg})$ are the proportions of observed positive (negative) comments whilst $pres_w$ and abs_w are the log-likelihood ratios of the events $(w \in r)$ and $(w \notin r)$, respectively.

While calculating those values for all the w ($w \in$ *Bag-of-Words*) words, it is possible to obtain an output such that reported in Table 1, where we have c_{pos}, c_{neg}, $pres_w$ and abs_w for each words in the considered *Bag-of-Words*.

Table 1 Threshold-based Naïve Bayes output

	w_1	w_2	w_3	w_4	w_5	...
$P(w_i\|c_{neg})$	0.011	0.026	0.002	0.003	0.003	...
$P(w_i\|c_{pos})$	0.007	0.075	0.005	0.012	0.001	...
$pres_{w_i}$	0.411	-1.077	-1.006	-1.272	1.423	...
abs_{w_i}	-0.004	0.052	0.003	0.008	-0.002	...

We have then used cross-validation to estimate a parameter τ such that: c is classified as "negative" if $LOR(c) > \tau$ or as "positive" if $LOR(c) \leq \tau$. Furthermore, in Table 2we have compared the performance of the proposed classifier (Tb-NB) with that of other well known competitors, in particular: Logistic Regression (LOG), Random Forest (RF), standard Naïve Bayes (NB E1071), Naïve Bayes using kernel estimated densities (NB KlaR), Decision Trees (CART), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

Table 2 Performance metrics on raw data using 5-fold cross validation

Classifier	ACC	Sensitivity	Fall-out	F1	MCC
Tb-NB	**0.911**	**0.929**	**0.117**	**0.926**	**0.813**
LOG	0.850	0.884	0.532	0.877	0.361
RF	0.811	0.873	0.591	0.849	0.303
NB(E1071)	0.806	0.804	0.389	0.834	0.390
NB(KLAR)	0.806	0.804	0.389	0.834	0.390
CART	0.768	0.842	0.587	0.815	0.272
LDA	0.764	0.860	0.641	0.816	0.246
SVM	0.793	0.930	0.290	0.771	0.621
average	0.805	0.893	0.377	0.810	0.508

Notes: ACC = Accuracy; F1 = F1-score; MCC = Matthews Correlation Coefficient

4 Conclusions

The ”versatile nature” of the objective score that the Threshold-based Naïve Bayes estimates improve the interpretability of the results. In fact, it is possible to observe in time different dimensions of services of an hotel or a country. Hence, evaluate the Customer Satisfaction of a certain area of interest like in Fig. 1 and Fig. 2.

References

1. Conversano, C., Romano, M., Mola, F.: Hotel search engine architecture based on online reviews' content, in: Smart Statistics for Smart Applications. Book of Short Papers SIS2019, G. Arbia, S. Peluso, A. Pini, and G. Rivellini (eds.), Pearson, Milan (IT), 213–218 (2019)
2. Esuli, A., Sebastiani, F.: SENTIWORDNET: A publicly available lexical resource for opinion mining, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijkn, and D. Tapias (eds.), European Language Resources Association (ELRA), Genoa (IT), 417–422 (2006)
3. Goldberg, Y.: . Neural Network Methods in Natural Language Processing. Synthesis Lectures on Human Language Technologies **10**(1), 1–309 (2017).
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), Curran Associates Inc., Lake Tahoe, Nevada (USA), 3111–3119 (2013)
5. Miller, G.A.: Wordnet: A lexical database for English. Communications of the ACM **38**(11), 39–41 (1995)
6. Romano, M., Frigau, L., Contu, G., Mola, F., Conversano, C.: Customer Satisfaction from Booking, in: Selected papers Conferenza GARR_18 Data (R)evolution. M. Mieli, and C. Volpe (eds.), Associazione Consortium GARR, Cagliari (IT), 111–118 (2018)

Threshold-based Naïve Bayes Classifier: Customer Satisfaction evaluation

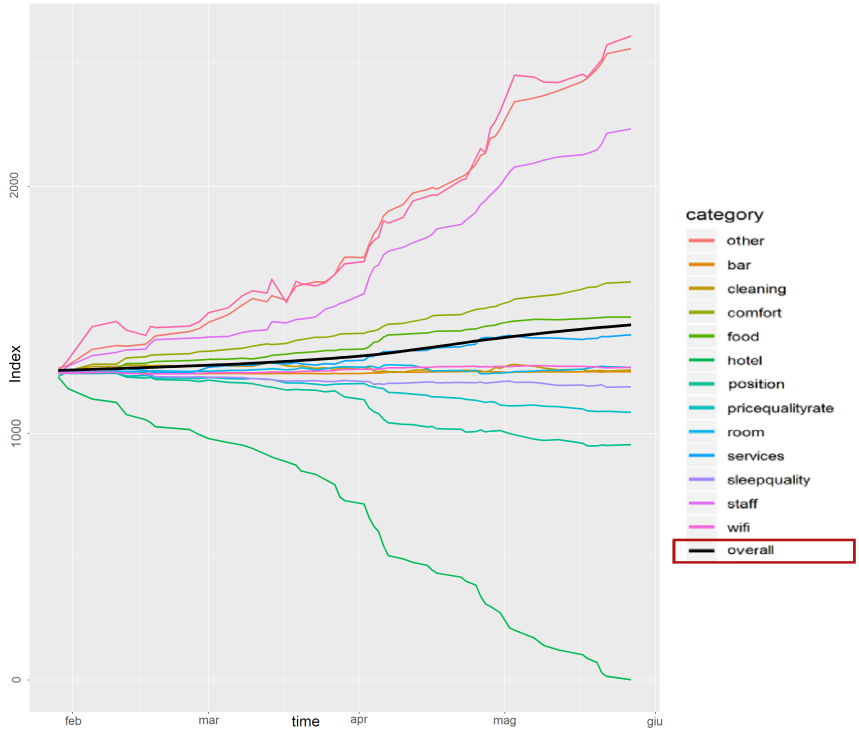


Fig. 1 Time-series for each category of services offered by a hotel scored 10/10 on Booking.com. From February 1, 2018 to May 27, 2018.

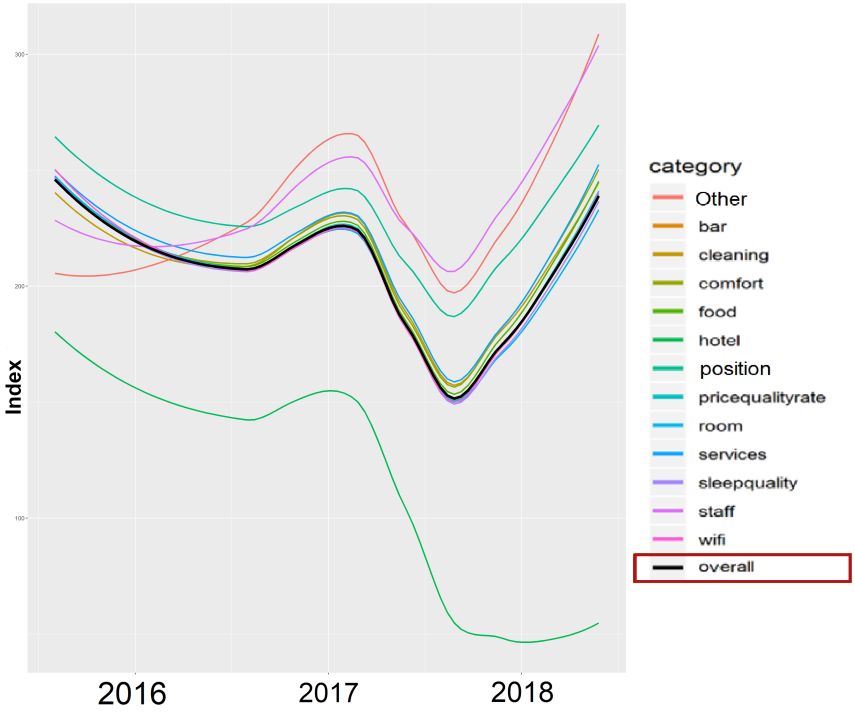


Fig. 2 Aggregated Booking.com data. Category scores observed in time (overall sentiment in black).

Session of free contributes SCL4 *–Society and Innovation*
Chair: Michelangelo Misuraca

Migrant Integration Policy Index (MIPEX): an analysis of countries via Gaussian mixture model-based clustering

Migrant Integration Policy Index (MIPEX): un'analisi internazionale attraverso un modello a mistura di gaussiane

Emiliano Seri, Leonardo Salvatore Alaimo, Enrico Di Bella, Rosanna Cataldo, Alfonso Piscitelli

Abstract In recent decades, there has been a growing research interest in comparative studies of migrant integration, assimilation and the evaluation of policies implemented for these purposes. With this aim, The Migrant Integration Policy Index (MIPEX), that measures policies to integrate migrants in 52 countries, has established itself as a solid reference on the subject over the years. In this work, we improve and facilitate the comparison between the treated countries by the application a Gaussian mixture model-based cluster analysis on the 8 MIPEX dimensions.

Abstract *Negli ultimi decenni, c'è stato un crescente interesse della ricerca per gli studi comparativi sull'integrazione e l'assimilazione dei migranti e la valutazione delle politiche attuate per questi scopi. Con questo obiettivo, il Migrant Integration Policy Index (MIPEX), che misura le politiche di integrazione dei migranti in 52 paesi, si è affermato negli anni come un solido riferimento sull'argomento. In questo lavoro, miglioriamo e facilitiamo il confronto tra i paesi trattati attraverso l'applicazione di una cluster analysis basata sul modello di mistura di Gaussiane sulle 8 dimensioni del MIPEX.*

Emiliano Seri
Department of Statistic, Sapienza University of Rome, e-mail: emiliano.seri@uniroma1.it.

Leonardo Salvatore Alaimo
Department of Social Sciences and Economics, Sapienza University of Rome, e-mail: leonardo.alaimo@uniroma1.it.

Rosanna Cataldo
Department of Social Sciences, University of Naples "Federico II", e-mail: rosanna.cataldo2@unina.it.

Enrico di Bella
Department of Political Sciences, University of Genoa, e-mail: enrico.dibella@unige.it.

Alfonso Piscitelli
Department of Agricultural Sciences, University of Naples "Federico II", e-mail: alfonso.piscitelli@unina.it.

Key words: Migrant Integration Policy Index, Model based clustering, Finite mixture models

1 Introduction

Immigration regulation and immigrant assimilation have been a salient political issue in all industrialised countries since many decades. The growing interest in comparative analyses of immigration has led to a variety of attempts to quantify immigration policies and to build indices. The study of these phenomena from a quantitative point of view is rather recent, due to the previous lack of data. Moreover, quantifying migrant integration is a difficult challenge, due to its complex nature and the lack of uniformity in the migration policy of many countries, which is based on multiple criteria. In a cross-country setting, to evaluate and compare what governments are doing to promote the integration of migrants, the Migrant Integration Policy Index (MIPEX) [1] has become a solid and useful tool . The project informs and engages key policy actors about how to use indicators to improve integration governance and policy effectiveness. For this purpose, the project identifies and measures integration policies and identifies the links between the latter, outcomes and public opinion, drawing on international scientific studies. Its aim is to measure policies that promote integration in both social and civic terms. In socio-economic terms, migrants must have equal opportunities to lead just as dignified, independent and active lives as the rest of the population. In civic terms, all residents can commit themselves to mutual rights and responsibilities on the basis of equality. The MIPEX includes 52 countries and collects data from 2007 to 2020, in order to provide a view of integration policies across a broad range of differing environments. It considers a system of 58 indicators (for more information, please consult [1]) covering 8 policy areas that have been designed to benchmark current laws and policies against the highest standards through consultations with top scholars and institutions[1] using and conducting comparative research in their area of expertise. The policy areas of integration covered by the MIPEX are the following:

- Labour Market Mobility
- Family Reunion
- Education
- Political Participation
- Long-term Residence
- Access to Nationality
- Anti-discrimination
- Health[2]

For each area, a synthetic measure (dimensional) is calculated as an arithmetic mean of the elementary indicators, i.e. those selected for measuring each policy area. Each dimensional synthetic indicator is bounded $[0, 100]$, in which the maximum of 100 is awarded when policies meet the highest standards for equal treatment. These

[1] The highest standards are drawn from Council of Europe Conventions, European Union Directives and international conventions (for more information see: http://mipex.eu/methodology)

[2] Health data are only available for years 2014 and 2019

values are chosen by experts from each country, by means of a questionnaire. Although not without its critics, MIPEX has become a reference for comparative studies on migrant integration over the last decade. The research question from which this paper starts is:

- *Given the complexity of the phenomenon under consideration, in order to improve the comparison between the countries considered, is it possible to identify homogeneous groups among them?*.

To answer this question, we applied a *Gaussian mixture model-based clustering* on the 52 considered countries for the eight dimensions of the MIPEX.

2 Methods

Given n independent observations identically distributed, $\mathbf{x} \equiv \{x_1, x_2, \ldots, x_n\}$, the distribution of each of them can be specified by a probability density function by means of a finite mixture model of G components as follows:

$$f(\mathbf{x}|\boldsymbol{\Psi}) = \sum_{k=1}^{G} \omega_k f_k(\mathbf{x}|\theta_\mathbf{k}) \tag{1}$$

where

$$\boldsymbol{\Psi} = \{\omega_1, \omega_2, \ldots, \omega_{G-1}; \theta_1, \ldots, \theta_G\}$$

are the parameters of the mixture model; $f_k(\mathbf{x_i}|\theta_\mathbf{k})$ is the k^{th} component density for observation x_i with parameter vector $\theta_\mathbf{k}$, $\{\omega_1, \omega_2, \ldots, \omega_{G-1}\}$ are the mixing weights, under the constraints:

$$\omega_j > 0 \quad \text{and} \quad \sum_{j=1}^{G} \omega_j = 1 \quad j = 1, \ldots, k$$

and G is the number of mixture components. Assuming that G is fixed, the mixture model parameters $\boldsymbol{\Psi}$ are usually unknown and must be estimated. Most applications assume that all component densities arise from the same parametric distribution family. A popular model is the Gaussian mixture model (GMM), which assumes a (multivariate) Gaussian distribution for each component ($f_k(\mathbf{x}|\theta_\mathbf{k}) \sim N(\mu_\mathbf{k}, \boldsymbol{\Sigma_k})$). Thus, a GMM is a weighted sum of G Gaussian component densities [2]:

$$f(\mathbf{x}|\boldsymbol{\Lambda}) = \sum_{i=1}^{G} \omega_i f_i(\mathbf{x}|\mu_i, \Sigma_i) \tag{2}$$

where $\boldsymbol{\Lambda} = \{\omega_i, \mu_i, \Sigma_i\}, i = 1 \ldots M$ are the model parameters, $\omega_i : i = 1, \ldots, M$ are the mixture weights, and $f_i(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i : i = 1, \ldots, M$, is the i^{th} Gaussian component density. A generic component density f_i is a D-variate Gaussian function of the form:

$$f_i(\mathbf{x}|\mu_i|\Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)'\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\} \tag{3}$$

where μ_i is the mean and Σ_i the covariance matrix. The complete GMM is parameterized by the mean vector, the covariance matrices and the mixture weights for all component densities. Model parameters are estimated by using the iterative Expectation-Maximization (EM) algorithm [3]. A GMM based clustering [4] allows to use a mixture of Gaussians to represent a population formed by G groups with weights $\omega_1, \omega_2 \ldots \omega_G$. An observation x can be classified into one of the G groups by computing the posterior probabilities:

$$p(g|x) = \frac{p_g f_g(x)}{\sum_h p_h f_h(x)}, \quad g = 1, \ldots, G \tag{4}$$

where $f(x)$ is the Gaussian density function. In GMM clustering approach, clusters are ellipsoidal, centered at the mean vector $\mu_\mathbf{i}$, and with other geometric features, such as volume, shape and orientation, determined by the covariance matrices $\boldsymbol{\Sigma_i}$. The choice of the optimal model and the optimal number of clusters is unsupervised and it is made according to the Bayesian information criterion (BIC) [6].

3 Application and results

As mentioned in Section 1, we proceed to analyse and cluster the data of the 8 MIPEX dimensional indicators. Figure 1 shows some useful descriptive statistics: above the main diagonal the Pearson's linear correlation coefficients (the correlation font is scaled by the size of the absolute correlation) and their significance level with confidence level $\alpha = 0.05$ are reported; on the main diagonal histograms and densities plot; below the main diagonal scatter plots and correlation ellipses[3]. The correlations show the relationships between the eight dimensions considered; the histograms and density lines give a graphical view of the form of the distributions of each indicator, while the scatterplots show graphically the relationships between each pair of indicators and their dispersion. We proceed to cluster the 52 countries considered by using a model-based approach via Gaussian mixture models. We choose a model-based approach to clustering, because we prefer an unsupervised approach, i.e. where choices such as the number of clusters, their shape and size and how clusters are assumed to differ, are made through inferential statistical methods (BIC coefficient). Among the possible distributional assumptions on the data, we focus on mixture of multivariate Gaussian densities for its ability to approximate the density function of any unknown distribution [7]. The clustering is computed via the *Mclust* package [5] of the **R** statistical software: the model selected is that with variable volume, equal shape and equal orientation (for details,

[3] The ellipse represents a level curve of the density of a bivariate normal with the matching correlation

Migrant Integration Policy Index

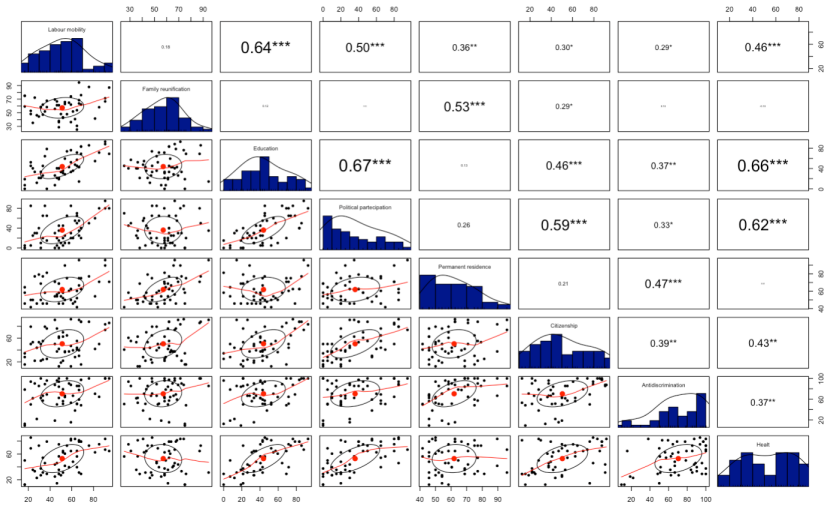


Fig. 1 Scatterplots, correlation ellipses, histograms, density plots and Pearson's linear correlation coefficients of the 8 dimensions of MIPEX.

please see: [5]) and 4 components (clusters). Table 1 reports the indicators' means of each component and the number of units.

Table 1 Means of components and number of units.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Labour market mobility	26.39	65.60	53.33	39.41
Family reunification	62.98	65.39	52.42	57.43
Education	9.50	76.08	42.12	25.17
Political participation	7.48	70.62	34.95	11.19
Permanent residence	51.04	68.51	59.57	64.27
Citizenship	35.93	78.80	48.80	28.35
Antidiscrimination	15.62	91.61	62.96	84.96
Healt	18.22	73.90	55.88	35.47
Number of units	4	12	25	11

Figure 2 shows the subdivision of the countries according to the cluster to which they belong. Cluster 1 comprises 4 countries: China, India, Indonesia and Russia. This cluster represents countries with the lowest level of integration policies for migrants. Cluster, including 12 countries, is the one of the "best integration", i.e. that present the higheest values in all the indicators. 25 countries are in the Cluster 3. This cluster group up the countries with average performances in all the indicators. The 11 remaining countries (Bulgaria, Croatia, Hungary, Latvia, Moldova, North Macedonia, Poland, Romania, Serbia, Slovakia, Slovenia) are classified in Cluster

4, characterised by low values in Political participation and Citizenship, but high in Permanent residence and the highest in Anti-discrimination.

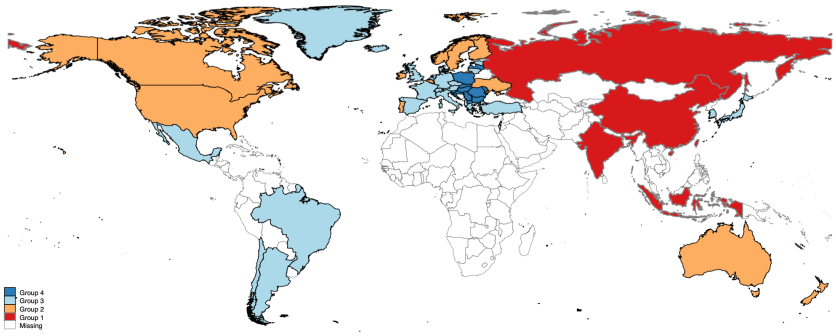


Fig. 2 MIPEX dimensional indices: clusters' composition of countries. Year 2019.

4 Conclusion

MIPEX aims to allow comparison of migration policies between countries. However, given the complex nature of the phenomenon analysed, the classification by means of Gaussian mixture model-based clustering, has made it possible to improve the reading of the results and therefore a better comparison and evaluation of the performance of the countries considered, for the 8 dimensions of MIPEX.

References

1. G. Solano, and T. Huddleston: "Migrant Integration Policy Index 2020.": 259. https://www.mipex.eu/. (2020).
2. D. A. Reynolds: "Gaussian mixture models." Encyclopedia of biometrics **741**: 659-663. (2009).
3. A. P. Dempster, N. M. Laird, and D. B. Rubin: "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society: Series B (Methodological) 39.1: 1-22. (1977).
4. G. J. McLachlan, and K. E. Basford: "Mixture models: Inference and applications to clustering". Vol. **38**. New York: M. Dekker, (1988).
5. L. Scrucca, et al.: "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." The R journal 8.1: 289. (2016).
6. G.Schwarz: "Estimating the dimension of a model." The annals of statistics: 461-464. (1978).
7. D. M. Titterington, et al.: "Statistical analysis of finite mixture distributions". Vol. **198**. John Wiley & Sons Incorporated, (1985).

The measure of the BES: a proposal for the aggregation of the indicator education and training

La misura del BES: una proposta per l'aggregazione dell' indicatore istruzione e della formazione

Luca Rossi[1] and Stefano Daddi[2]

Abstract Measures of economic progress, well-being and social well-being are widely studied in literature, but various factor make this operation difficult. In this article, we propose a synthetic index for dimension education and training, based on P_2 distance defined by Pena, to rank the 20 Italian Regions. For this purpose, we used the five simple indicators inserted in the BES 2016 report by National Institute of Statistics (ISTAT). The proposed methodology for the construction of index solves problems of aggregation of variables expressed in different units and arbitrary weights.

Abstract *Le misure del progresso economico, del benessere e del benessere sociale sono ampiamente studiate in letteratura, ma vari fattori rendono difficile questa operazione. In questo articolo proponiamo un indice sintetico per la dimensione istruzione e formazione, basato sulla distanza P_2 definita da Pena, per classificare le 20 Regioni italiane. A tal fine abbiamo utilizzato i cinque semplici indicatori inseriti nel rapporto BES 2016 dell'Istituto Nazionale di Statistica (ISTAT). La metodologia proposta per la costruzione dell'indice risolve problemi di aggregazione: variabili espresse con unità di misura differenti o pesi arbitrari.*

Key words: BES, synthetic indicator, Pena P_2 distance, education and training

[1] Luca Rossi, University "Niccolò Cusano"; email: luca.rossi@unicusano.it

[2] Stefano Daddi, ISTAT; email: daddi@istat.it

1 Introduction

Measures of economic progress, well-being and social well-being are used by governments as cornerstones for designing public policies (Jayawickreme et al. 2012; Layard 2011; Sachs 2012). They accurately highlight the changes that occur both in individual living standards (Helliwell et al. 2012) and in the complex economic growth of a State (Diener et al., 2009). This information is important in countries like Italy divided into political entities (20 Regions), each authorized to protect the interests of their respective communities. At the national level, it is therefore essential to understand the role that each Region plays in defining the economic performance and social progress of a State.

In Italy, an indicator called "Benessere Equo e Sostenibile" (BES) is used to evaluate the progress of society not only from an economic, but also from a social and environmental point of view. Since 2010, traditional economic indicators, first of all the Gross domestic product (GDP), have been integrated with measures on the quality of life of people and the environment. Starting from 2016, indicators and analyses on well-being are flanked by indicators for monitoring the objectives of the 2030 Agenda on sustainable development: The United Nations Sustainable Development Goals (SDGs). In particular, the Italian Government with the Committee for BES indicators has enforced the analysis of 12 dimensions, (each consisting of a different number of simple indicators) deemed necessary for an optimal measurement of the BES. Therefore, the final goal of our research will be to determine a composite index that synthesizes the 152 simple indicators inserted in the BES report by National Institute of Statistics.

The first step will necessarily be to synthesize the elementary indicators within each dimension. In this work we propose a statistical methodology to obtain a composite indicator for "education and training" dimension that allows a classification of the Italian territorial micro-areas providing a clearer view of the differences in the education and training sector.

2 Methodology

According with McGregor (2015) and Bache (2019), the well-being indicators highlight the role of well-being in the formulation of national and international policies. Such indicators, providing a more objective, solid and reliable information base, should be measures of government performance.

From a technical point of view, there are various methods that allow to aggregate and normalise a set of input variables to create valid and robust composite indicators. The purpose of these methods is to simplify a multidimensional analysis according to a formative or reflective measurement model, where the elementary indicators are respectively causes or effects of the latent variable (Michalos, 2014). Principal component analysis (PCA) and Categorical Principal Components Analysis (CATPCA) for metric variables and nominal, ordinal and continuous variables respectively are methods widely used to reduce the multidimensionality of the variables considered in many sectors including the analysis of economic development

or quality of life of a country. However, when the goal is to build a composite indicator, PCA or CATPCA, nevertheless used, are inappropriate mainly because these methods ignore the polarities, the assignation of arbitrary weighting within the synthetic indicator (Jollife et al., 2016). Another fundamental aspect, to build a reliable index, is the selection of the measurement model. This defines whether the relationships between the latent variable (the composite phenomena) and the elementary indicators are determined through a formative or reflective form (Mazziotta et al., 2019). Sustainable well-being measures are based on objective indicators such as health, services or environmental quality. This implies that the models are formative because the latent measures are derived by elementary indicators (Maggino et al., 2012).

In line with the above, the main objective of this paper is to construct a synthetic indicator for dimension education and training using a methodology that solves the problems previously expressed. To achieve our objective, we have designed an indicator focused on the distance P_2 (DP_2), defined by Pena (1977). This methodology is widely used, in various areas, for building indicators that summarize a set of elementary variables that describe the different aspects of the phenomenon under test. The major research areas are those related social well-being (Somarriba et al., 2016; Cuenca et al., 2010). and quality of life (Rodríguez et a., 2018; Martín et al., 2019). According with Somarriba and Pena (2009), this metrics has various properties such as non-negativity, commutativity, triangular inequality, existence, determination, monotony, transitivity, neutrality, invariance to a change of origin and/or scale and solves problems such as aggregation of variables expressed in different units, arbitrary weights and information duplicity.

The DP2 indicator is defined as:

$$DP_2 = \sum_{j=1}^{n} \frac{d_{ij}}{\sigma_j} * \left(1 - R^2_{j,j-1,\dots,1}\right) \quad \forall\, i \in [1,m]$$

where $j = 1,2,\dots.n$ are the input variables and $i = 1,2,\dots.m$ are Regions.
$d_{ij} = \left|x_{ij} - x_{jmin}\right|$ is the difference between the value of $j-th$ variable in the $i-th$ Region and the minimum of $j-th$ variable in all the Regions.
σ_j is the standard deviation of the $j-th$ variable and is used as scaling factor and allows to solve the problem of heterogeneity of unit measures
$R^2_{j,j-1,\dots,1}$ is the coefficient of determination in the regression of x_j over x_{j-1},…. x_1As defined by Pena, it is a "correction factor" which measures the variation for each variable elucidated by the linear regression relative to preceding variables eliminating, in this way, redundant information from the variables already in the synthetic index.

3 Results

The aim of this paper is to construct a synthetic indicator for dimension "education and training" (E&T-P_2). To achieve our scope, we used the 5 simple indicators inserted in the "education and training" dimension (table 1) inserted in BES 2016

report by National Institute of Statistics (ISTAT), considering the 20 Regions. In line with the above, the index has been constructed using the P_2 distance defined by Pena.

Table 1: education and training structure (Source: The authors)

Variables included in education and training BES 2016 Report	Correlation coefficient	Correction factors
Participation in continuing education	0.9748	1.0000
People with at least a diploma (25-64 years)	0.7790	0.4817
Graduates and other tertiary qualifications (30-34 years)	0.6833	0.2746
Early exit from the education and training system	-0.5183	0.3362
Participation in the school system of 4-5 year olds	-0.5204	0.3087

Table 1 highlights the entry order of variables by correlation coefficient and the correction factors representing the contribute that each simple indicator gives to E&T-P_2 index. The results indicate that all partial indicators have appreciable correction factors and this implies that contribute to E&T-P_2 with useful and not duplicated information. The partial indicator that has highest positive correlation coefficient is "Participation in continuing education". The result is not surprising given that a continuous training course is aimed at increasing the skills necessary to increase people's employability. This first variable has a correction factor of 100%. This means that 100% of the information of this indicator is introduced in the measurement of E&T. The positive coefficient of variable "Graduates and other tertiary qualifications (30-34 years)" is obvious because those with a degree have a diploma too but only the 27% the information for this variable is incorporated into E&T-P_2. The strong negative correlation coefficient of "Early exit from the education and training system" reflects what has been highlighted: low level of education means fewer opportunities to find work (Gesthuizen et al., 2014). The negative correlation coefficient of variable "Participation in the school system of 4-5 year olds" implies Italy is not yet able to offer all young people the possibility of adequate education. The findings highlights that education, training and skills level influence people's well-being and open opportunities that would be otherwise precluded.

The second step of this work has been to determine the situation of the 20 Italian Regions in the considered domain. In table 2 is showed the classification resulting from the synthetic index E&T-P_2. In 2015 the Region with the highest levels of education and training is Trentino-Alto Adige with a distance of 7.4166 points from the imaginary baseline Region. Puglia is the Region that is least able to provide adequate education. The top ten of Regions resulting from synthetic index is made up of those countries that offer healthier lifestyles and have more opportunities to find work in less risky environments.

Table 2: Classification of Regions index E&T-P2 (Source: The authors)

Regions	E&T-P2
Abruzzo	4.413530
Basilicata	3.080197
Calabria	3.292989

The measure of the BES: a proposal for the aggregation of the indicator education and training

Campania	2.593854
Emilia-Romagna	5.029744
Friuli-Venezia Giulia	5.884764
Lazio	4.967399
Liguria	4.278304
Lombardia	4.549199
Marche	4.345765
Molise	4.227538
Piemonte	4.011572
Puglia	2.346952
Sardegna	4.226782
Sicilia	2.440774
Toscana	5.411202
Trentino-Alto Adige	7.416591
Umbria	5.320694
Valle d'Aosta	4.181322

To verify the goodness of the proposed index we compared the obtained classification of Regions with that obtained by ISTAT applying Mazziotta-Pareto index (MPI) and described in BES 2016 report (page 49). The results obtained with both indices seem to interpret the phenomenon considered in the same way, providing both the same classification. The method P_2 distance has the property of neutrality and this means that the weighting is not determined in advance, as may occur in an aggregation method such as MPI, but as a result of the calculation procedure.

References

1. Bache, I.: How does evidence matter? Understanding 'what works' for wellbeing. Social Indicators Research **142**(3), 1153–1173 (2019)
2. Cuenca, E., Rodríguez, J. A., & Navarro, M.: The features of development in the Pacific countries of the African, Caribbean and Pacific group. Social Indicators Research **99**, 469–485 (2010)
3. Diener, E., Lucas, R., Schimmack, U., & Helliwell, J.: Well-being for public policy. Oxford, UK: Oxford University Press. (2009)
4. Gesthuizen, M., Solga, H.: Is the labor market vulnerability of less-educated men really about job competition? New insights from the United States. J Labour Market Res **47**, 205–221 (2014)
5. Helliwell, J. F., Layard, R., & Sachs, J.: Some policy implications. In J. F. Helliwell, R. Layard, & J.Sachs (Eds.), World happiness report (pp. 90–107) (2012). New York, NY: The Earth Institute, Columbia University
6. ISTAT. BES 2016 Report
7. Jayawickreme, E., Forgeard, M. J., & Seligman, M. E.: The engine of well-being. Review of General Psychology **16**(4), 327–342 (2012)
8. Jolliffe IT, Cadima J.: Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci **374**(2065) (2016).
9. Layard, R.: Happiness: Lessons from a new science (2nd ed.) (2011). New York, NY: Penguin Press.
10. Maggino, F., & Zumbo, B. D.: Measuring the quality of life and the construction of social indicators. In K. C. Land, A. C. Michalos, & M. J. Sirgy (Eds.), Handbook of social indicators and quality-of-life research (pp. 201–238) (2012). Dordrecht: Springer
11. Martín, J. M., Salinas, J. A., & Rodríguez, J. A.: Comprehensive evaluation of the tourism seasonality using a synthetic DP2 indicator. Tourism Geographies **21**(2), 284–305 (2019)
12. Mazziotta, M., & Pareto, A.: Use and misuse of PCA for measuring well-being. Social Indicators Research **142**(2), 451–476 (2019)

13. McGregor, J. A.: Global initiatives in measuring human wellbeing: Convergence and diference. CWiPP Working Paper **2**, (2015). Shefeld: Centre for Wellbeing in Public Policy, University of Shefeld
14. Michalos, A. C.: Encyclopedia of quality of life and well-being research. (2014). Berlin: Springer
15. Pena, J. B.: Problemas de la medición del bienestar y conceptos afines (una aplicación del caso español). (1977). Madrid: INE
16. Pena, J. B.: La medición del bienestar social: Una revisión crítica. Estudios de Economía Aplicada **27**(2), 299–324 (2009).
17. Rodríguez, J. A., Jiménez, J. D., Martín, J. M., & Salinas, J. A.: Crisis in the horn of Africa: Measurement of Progress towards millennium development goals. Social Indicators Research **135**, 499–514 (2018)
18. Sachs, J. Introduction. In J. F. Helliwell, R. Layard, & J.: Sachs (Eds.), World happiness report. New York, NY: Columbia University (2012)
19. Somarriba, N., & Pena, B.: Synthetic indicators of quality of life in Europe. Social Indicators Research **96**, 115–133 (2009)
20. Somarriba, N., & Zarzosa, P.: Quality of life in Latin America: A proposal for a synthetic indicator. In indicators of quality of life in Latin America. Social Indicators Research Series **62**, (2016)

Monitoring consumer sentiment using control charts

Monitorare le opinioni dei consumatori con le carte di controllo

Marina Marino, Rocco Mazza, Michelangelo Misuraca and Agostino Stavolo

Abstract Product and service reviews written by consumers as free texts are more and more a fundamental data source to evaluate customer satisfaction. In this work, we propose a strategy to draw a control chart based on the semantic orientation of the opinions conveyed in the reviews. This tool can be used by firms as a visual analytic to monitor consumer experiences and detect which items have to be pushed or withdrawn. A case study based on a set of Amazon reviews is briefly discussed.

Abstract *Le recensioni di prodotti e servizi scritte dai consumatori come testi liberi sono sempre più una fondamentale fonte di dati per valutare la soddisfazione dei clienti. In questo lavoro si propone una strategia per disegnare una carta di controllo basata sull'orientamento semantico delle opinioni contenute nelle recensioni. Tale strumento può essere usato dalle aziende per monitorare le esperienze dei consumatori e individuare quali articoli spingere o eliminare. Un caso studio su recensioni di Amazon è qui brevemente discusso.*

Key words: opinion mining, polarity score, statistical process control

1 Introduction

The technological progresses and the increasing diffusion of Web 2.0 during the last few years changed the way people communicate feelings, opinions and experiences about each facet of everyday-life. More and more, social networks, online forums and review sites and weblogs, are used to share personal ideas and viewpoints and exchange information. This new habit impacted also on purchase intention, since people often decide to buy a product or a service after evaluating the positive and

Marina Marino, Rocco Mazza, Agostino Stavolo
University of Naples Federico II, email: marina.marino@unina.it; rocco.mazza@unina.it; agostino.stavolo@unina.it

Michelangelo Misuraca
University of Calabria, email: michelangelo.misuraca@unical.it

negative statements made by those that already bought the same product or service under consideration. The so-called *electronic word-of-mouth* became then a primary source of information at an individual level but also at a firm level [2,7].

Analysing people comments and reviews is essential for any kind of business, since dissatisfied consumers are potentially dangerous to the firms, triggering a vicious circle of bad reputation [19] and considerable financial losses [16]. Thus, in a managerial perspective, it is necessary to monitor what consumers say about their consumption experiences and feelings [17]. Since comments and reviews are written as free texts, information concerning the opinions are encoded in a form difficult to process automatically. In a text mining framework, it is possible to pre-treat the textual body in order to transform the unstructured data into structured data, and then perform statistical analyses to extract and manage the underlying knowledge base. Dealing with opinions, it is particularly interesting to consider the semantic orientation of the texts, expressing in a numerical form the so-called *sentiment*. Several alternative approaches have been proposed concerning how to calculate a score expressing the negative/positive orientation embodied in the texts and use these scores in an opinion mining strategy [6,15]. Nevertheless, the visualisation of sentiment is still an open research topic, with very few innovative contributions [13], and more oriented towards topic extraction [21,26] or classification [12]. Aiming at monitoring the orientation and the intensity of consumer sentiment, an interesting framework is offered by Statistical Process Control (SPC) techniques and by control charts in particular. Control charts have been widely used in industry as a tool to monitor product features [23]. Typically, they are implemented to control ongoing processes, predict the expected range of outcomes from a process, determining whether a process is stable in statistical control, analysing patterns of process variation from non-routine events or built-in events [24]. The sentiment emerging from the reviews written by consumers for a given product/service, can be seen as a quantitative characteristic of a process involving their purchases as well as their feelings about product quality, usability and functionality. For this reason, trying to build an easily readable visual analytic, here we propose a strategy to monitor the sentiment by using control charts. After briefly reviewing the reference literature and showing in details the adopted methodology, we present the preliminary results of a case study based on a set of reviews concerning the purchase of cell phones and related accessories on Amazon.

2 A brief literature review

SPC methods were initially developed to monitor the quality of products, later extending their use also to services. Control charts are generally classified into two groups. If the quality characteristic is measured on a continuous scale, we have control charts for variables. When instead the quality characteristic is classified as conforming or not-conforming considering the possess of given characteristics, we have control charts for attributes. Control charts have been applied in a large number of diverse areas. Control charts have been proposed, for example, in livestock farming to monitor

and manage animal production systems [3]. In a health-care domain, control charts have been used for public health surveillance [4,22] and surgery quality [8]. In survey design, some authors proposed the use of control charts to check data collection quality [5,9]. In a business domain, aside from manufacturing and product quality control, SPC and control charts have been also used to monitor customer satisfaction and perceived quality [10]. Customer survey data are typically collected considering both qualitative and quantitative characteristics.

As stated above, the use of textual data can offer in this field valuable insights by analysing directly what consumers say or think about a given product or service. In a more general framework concerning the statistical analysis of document collections, [1] suggested the application of control charts to evaluate the solution of a text mining strategy based on a Latent Semantic Analysis. The authors, in particular, evaluated topics explaining the sources of customer satisfaction or dissatisfaction, considering each case of non-conformity as an element requiring an in-depth investigation. This work represented an important crossroads in the reference literature body, since it focuses on unstructured textual feedback and establishes a mechanism for transitioning the textual analysis into an actionable SPC. The assumption underlying this approach is that "out-of-control" comments are negative and express a complain. The analysis of negative and positive comments is the core of opinion mining. [14] extended the SPC approach to the sentiment analysis of consumer reviews, detecting shifts in topic-sentiment combinations. Similarly, [11] proposed a sentiment-based SPC to systematically identify critical complaints within customer review data. Also in the latter case, the analysis specifically focused on negative comments, adopting a complaints index based on the measure used by [25] for service quality control. Here we propose an SPC strategy to monitor customer satisfaction at a product level, monitoring the polarity score attributed to consumer reviews written in natural language. The rationale is to check at the same time if a product or a service has a negative or positive opinion, offering to firms with a huge product/service catalogue a visual tool able to easily identify which items is necessary to push or withdraw.

3 A sentiment-based control chart

The analytical strategy based on the SPC of consumer reviews' sentiment is carried on in three different steps:

1. consumer reviews are retrieved and stored in a repository, together with available metadata (e.g., item ID, item rating, publishing date), then the textual body is pre-treated by cleansing and normalising the embodied terms;
2. for each comment in the repository, a polarity score is calculated to express the negativity/positivity of the underlying opinion about the item;
3. a control chart is drawn to detect the most relevant variations within the sentiment distribution over the item under control.

In the first step, after retrieving and storing the *n* reviews under investigation, texts are "tokenised" to obtain sets of distinct strings (namely, *tokens*) representing the

different terms embodied in the texts, separated by blanks and punctuation marks. All non-alphabetic characters and symbols – like numbers or emoticons – are removed, to consider in the following steps only content-bearing terms. Tokens are then reduced to lowercase. At the end of the pre-process, each text can be encoded as a *p*-dimensional vector, where *p* is the total number of terms listed in the collection.

In the second step, a polarity score is calculated for each review. In particular, each text is split into its composing sentences to take into account the sentiment associated with the different aspects concerning the item review made by the consumers. Given a review d_i, with $i=[1,n]$, its q_i sentences $\{s_{i1}, ..., s_{ik}, ..., s_{iq_i}\}$ are identified by considering as separators strong punctuation marks like full stops, question marks and exclamation marks. The *k*-th sentence is represented as a sequence of its p_k terms $\{w_{ik1}, ..., w_{ikj}, ..., w_{ikp_k}\}$, preserving the order of terms into the sentence. Each term w_{ikj} in the *k*-th sentence of the *i*-th review is compared with a dictionary of polarised terms, giving a score $PS_{w_{ikj}}$ of -1 to negative terms and a score of +1 to positive terms, respectively. Terms not listed in the dictionary are considered neutrals and scored with a 0. The polarity of each term is weighted considering negator terms (e.g., *never*, *none*), amplifier and de-amplifier terms (e.g., *very*, *few*), adversative and contrasting terms (e.g., *but*, *however*). This weighting scheme allows emphasising or dampening the negativity or positivity of each term, leading to a more effective measure of the polarity [20]. The $PS_{s_{ik}}$ polarity score of each sentence is obtained as the sum of its weighted term scores $PS_{w_{ikj}}$, on the square-root of the sentence length:

$$PS_{s_{ik}} = \frac{\sum_{j=1}^{p_k} PS_{w_{ikj}}}{\sqrt{p_k}} \tag{1}$$

Since we are interested in obtaining a polarity score at a review level, we calculate an overall score PS_{d_i} for each text by a down-weighted zeros average of sentence polarities, giving a minor weight to sentences conveying a neutral sentiment:

$$PS_{d_i} = \frac{\sum_{k=1}^{q_i} PS_{s_{ik}}}{q_i^{NEG}+q_i^{POS}+\sqrt{\log(1+q_i^{NEU})}} \tag{2}$$

where q_i^{NEG}, q_i^{POS} and q_i^{NEU} are the number of sentences in d_i with a negative, positive, or neutral polarity, respectively. The score PS_{d_i} assumes values in a $]-\infty,+\infty[$ interval.

In the third step of the strategy, the process to be statistically controlled is defined and then the resulting control chart is drawn and analysed. The logic beneath the process is controlling the distribution of sentiments expressed by consumers in the reviews of the monitored items and detecting out-of-control points. The distance from the central line of the chart is measured in terms of standard deviation, as the maximum and minimum extent of the divergence to label a point as out-of-control. The use of 3σ to plot upper and lower warnings is well established in the reference

literature [18] when SPC is applied to industrial mass production. Since we refer to the description of a process of opinion creation within a community of consumers, the sensitivity parameter can be set on values lower than 3.

4 Some preliminary results

We collected the data from an online repository[1] including 142.8 million reviews published on Amazon between May 1996 and July 2018. The dataset includes reviews (ratings, texts, helpfulness votes), product metadata (descriptions, category, price, brand, and image features), and links ("also viewed"/"also bought" images). In particular, we focused only on reviews about cell phones and accessories (3.4 mln), and considered the products with at least10 reviews and with verified purchases (842,000). Here we considered only the textual information, the date of the review and the so-called ASIN, a code used to identify uniquely a given product.

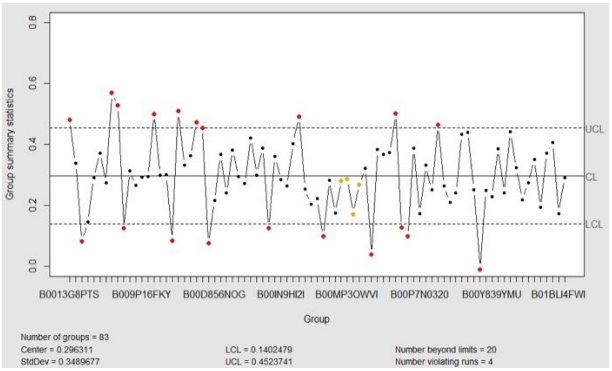


Figure 1: Sentiment-based control chart on products

After calculating the polarity for each comment, we generated a sentiment matrix in which for each ASIN (rows) we selected the 20 most recent comments (columns). This choice aims at avoiding variations in the process caused by modifications or updating of the products made by the sellers. The control line was defined as the average sentiment on each ASIN and the warnings considered a sensitivity parameter equal to 2. The chart in Fig. 1 highlights possible anomalies on the distribution. Products with a sentiment above the upper limit line can be described as strongly recommended products. Basically, the buying community highly recommended its purchase. On the other side, products below the lower limit line can be defined as disappointing products. Although these products do not showed an overall negative sentiment, they are the ones that deviate the most from the control limit, so they deserve more attention. Results will be discussed more in details elsewhere.

[1] See https://nijianmo.github.io/amazon

References

1. Ashton, T., Evangelopoulos, N., Prybutok, V.: Quantitative quality control from qualitative data: control charts with latent semantic analysis. Qual. Quant. **49**, 1081–1099. (2015)
2. Cheung, C.M.K., Thadani, D.R.: The impact of electronic word-of-mouth communication: A literature analysis and integrative model. Decis. Support Syst. **54**, 461–470 (2012)
3. De Vries, A., Reneau, J.K.: Application of statistical process control charts to monitor changes in animal production systems. J. Anim. Sci. **88**, E11–E24 (2010)
4. Finison, L.J., Finison, K.S, Bliersbacl, C.M.: Use of control chart to improve healthcare quality, J. Healthc. Qual. **15**, 9–23 (1993)
5. Gonzalez, Y., Oliver, B.: Producing Control Charts to Monitor Response Rates for Business Surveys in the Economic Directorate of the U.S. Census Bureau. In: 2012 FCSM Research & Policy Conference, pp. 1–12. https://nces.ed.gov/FCSM/pdf/Gonzalez_2012FCSM_V-C.pdf (2012)
6. Hemmatian, F., Sohrabi, M.K.: A survey on classification techniques for opinion mining and sentiment analysis. Artif. Intell. Rev. **52**, 1495–1545 (2019)
7. Ismagilova, E., Slade, E.L., Rana, N.P., Dwivedi, Y.K.: The Effect of Electronic Word of Mouth Communications on Intention to Buy: A Meta-Analysis. Inf. Syst. Front. **22**, 1203–1226 (2020)
8. Jaffray, B.: Am I out of control? The application of statistical process control charts to children's surgery. J. Pediatr. Surg. **55**, 1691–1698 (2020)
9. Jin, J., Loosveldt, G.: Assessing Response Quality by Using Multivariate Control Charts for Numerical and Categorical Response Quality Indicators. J. Surv. Stat. Methodol. **9**, 674–700 (2021)
10. Kenett, R.S., Deldossi, L., Zappa, D.: Quality standards and control charts applied to customer surveys. In: Kenett, R.S., Salini, S. (eds) Modern Analysis of Customer Satisfaction Surveys, pp. 413–438. Wiley, Chichester (2012)
11. Kim, J., Lim, C.: Customer complaints monitoring with customer review data analytics. An integrated method of sentiment and statistical process control analyses. Adv. Eng. Inf. **49**, 101304 (2021)
12. Kim, K., Lee, J.: Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. Pattern Recogn. **47**, 758–768 (2014)
13. Kurcher, K., Paradis, C., Keren, A.: The state of the art in sentiment visualization. Comput. Graph. Forum. **37**, 71–96 (2018)
14. Liang, Q., Wang, K.: Monitoring of user-generated reviews via a sequential reverse joint sentiment-topic model. Qual. Reliab. Eng. Int. **35**, 1180–1199 (2019)
15. Liu, B., Zhang, L.: A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer, Boston (2012)
16. Luo, X.: Quantifying the Long-Term Impact of Negative Word of Mouth on Cash Flows and Stock Prices. Market. Sci. 28, 148–165 (2008)
17. Nambisan, P., Watt, J.H.: Managing customer experiences in online product communities. J. Bus. Res. **64**, 889–895 (2011)
18. Nelson, L.S.: When should the limits on a Shewhart control chart be other than a center line ±3-sigma? J. Qual. Technol. **35**, 424–425 (2003)
19. Shirdastian, H., Laroche, M., Richard M.-O.: Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. Int. J. Inform. Manag. **48**, 291–307 (2019)
20. Vechtomova, O.: Disambiguating context-dependent polarity of words: An information retrieval approach. Inf. Process. Manag. **53**, 1062–1079 (2017)
21. Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., Zhang, K.: SentiView: Sentiment Analysis and Visualization for Internet Popular Topics. IEEE Trans. Hum. Mach. Syst. **43**, 620–630 (2013)
22. Woodall, W.H.: The Use of Control Charts in Health-Care and Public-Health Surveillance. J. Qual. Technol. **38**, 89–104 (2006)
23. Woodall, W.H., Spitzner, D.J., Montgomery, D.C., Gupta, S.: Using Control Charts to Monitor Process and Product Quality Profiles. J. Qual. Tech. **36**, 309–320 (2004)
24. Xie M., Goh T.N., Kuralmani V.: Statistical models and control charts for high-quality processes. Springer, New York (2002)
25. Yang, H.H., Chen, K.S.: A performance index approach to managing service quality. Manag. Serv. Qual. **10**, 273–278 (2000)
26. Zhao, Y., Qin, B., Liu, T., Tang, D.: Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog. Multimed. Tools Appl. **75**, 8843–8860 (2016)

Session of solicited contributes SS3 – *Statistical methods for the assessment of student careers in higher education*
Organizer and Chair: Maria Prosperina Vitale

Predicting university students' churn risk

La previsione del rischio di abbandono dell'Ateneo da parte dei laureati triennali in fase di iscrizione alla laurea magistrale

Michele La Rocca, Marcella Niglio and Marialuisa Restaino

Abstract The aim of the paper is to estimate university students' churn prediction, where university churn is defined as the abandonment of students that, after earning their first level graduation at a university, decide to continue their studies in other universities. The results can be used to identify the profiles of the students that are willing to enroll in a master program from the same university. We merge more datasets from different sources to obtain the set of possible risk factors that might affect the churn risk. Since the number of potential factors is particularly high, we adopt a penalized variable selection method to identify the most relevant covariates and improve the final model interpretability.

Abstract *Lo scopo del presente lavoro è la previsione del rischio di abbandono dell'Ateneo nel quale è stata conseguita la laurea triennale da parte di studenti che si iscrivono alla laurea magistrale. In particolare l'analisi è condotta per valutare il profilo degli studenti magistrali che decidono di continuare a studiare nell'Ateneo nel quale hanno conseguito la laurea triennale. L'approccio utilizzato è basato su metodi penalizzati di selezione di variabili al fine di selezionare, in un ampio dataset, le covariate più rilevanti che garantiscano un'adeguata interpretazione del modello stimato.*

Key words: Students' churn, variable selection, AlmaLaurea, glm models

1 Introduction

Churn analysis plays a crucial role in profit companies to understand why customers have stopped using products or services. Analyzing the churn of a company doesn't only mean knowing what the churn rate is. Most importantly, it is about figuring out why customers are churning at a given rate and how to fix the problem. A distinction is usually made between *involuntary churn*, which occurs when the company terminates the customers' contract, or *voluntary churn*, which occurs when the customer decides to take their business elsewhere.

In this paper, this general framework is adapted to the problem of students' mobility after the first-level degree, also called *second level mobility* in [1].

Michele La Rocca, Marcella Niglio and Marialuisa Restaino
University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy, e-mail: [larocca, mniglio, mlrestaino]@unisa.it

The study of university students careers often aims to evaluate the elements that contribute to the risk of leaving university without reaching any degree (shortly said *university dropout*). In other cases, it also considers the mobility among courses of study. In this latter case, the risk of churn is related to the abandonment of a course and the subsequent enrollment elsewhere (*dismissed student*). Even if these phenomena are of great interest and strongly affect the Italian university system, we change the point of view here. By *university churn*, we mean the abandonment of students that, after earning their first level graduation (*bachelor degree*) at the University of Salerno, decide to continue their studies in other universities. Specifically, we are interested in a group of students who might choose to continue their careers in the same university where they have earned the first level degree or in other universities. Following [4, 3], a student that decides to churn is defined *dismissed student*.

The contributions given in the literature have differently evaluated the determinants of the first and second level students' mobility: the individual skills, the family background, the characteristics of the region of origin, the location of the universities, the job opportunities etc. (see among the others [6] , [2], [1] and the references therein).

Here, our aim is to define students' profiles with high churn risk, depending on a set of risk factors. In particular, the university churn is related to the choice of students to continue their study in the same university where they graduated in the first level course or to change university. Thus, the variable of interest is binary and takes value 1 if students are enrolled at the University of Salerno for the master program and value 0 if they go to other universities. The risk factors that can influence the churn risk are related to the educational and job paths and socio-demographic characteristics. Therefore, the obtained profiles could be used to implement appropriate university policies to reduce the churn risk.

The paper is organized as follows. Section 2 describes the merging process of the used data sets. Section 3 discusses the methodological proposal for churn risk estimation. Finally, Section 4 reports the empirical evidences and some concluding remarks.

2 Data set

The complexity of the factors contributing to the university churn under analysis has required the analysis of large datasets from different sources and, consequently, the need for careful data wrangling steps, including data integration, data linking and data transformation.

In particular, we have collected three different databases: i) the first extracted from the data warehouse of the University of Salerno (ESSE3); ii) the second from the AlmaLaurea Consortium (www.almalaurea.it) on the graduates' profile survey, whose annual data are delivered to all joined universities; and iii) the third extracted from the graduates' employment status survey (one year after the graduation), annually performed by AlmaLaurea that delivers the micro-data to the universities.

The ESSE3 data set is obtained after merging information on students enrollment, exams and graduation for all years under analysis. The two AlmaLaurea datasets are characterized by a large number of variables (i.e. in the last survey on the graduates' profile there are 145 variables, whereas the dataset of the graduates' employment status counts 159 variables). It highlights not only the complexity of the final merged data set, but also the need to properly select the potentially relevant variables. Thus, a first screening of the covariates has been performed to reduce the number of risk factors by taking into account the aim of the study.

At this stage of the research, we have considered the graduated students of the first level courses of study of the Department of Economics & Statistics (University of Salerno) given by the course in Business Administration (BA), Economics (E) and Administration and Organization (A&O). The analysis covers 1,543 students (BA=697; E=654; A&O=192) that have started a master program at the University of Salerno (1,036 students) or elsewhere (507 students). The analysis covers eight years (2013-2020) and the merging key used to rebuild the graduated students record, among the three databases, has been students' identification number.

In more detail, given the aim of the research, we have considered for the eight years the variables described in Table 1, where each of them has been classified considering the information included. Also a short description and the type of data are given in the following two columns. In particular, starting from some information related to the high school, we have considered some factors related to their first level university experience, job position and socio-demographic variables.

3 Model strategy

The binary nature of the dependent variable of interest and the large number of variables that can affect the churn risk in Table 1, have lead to choose a Logit regression model whose parameters have been estimated using penalized estimation techniques.

Let Y be a Bernoulli random variable. The conditional probability that $Y = 1$, given a vector of p covariates $\mathbf{x}' = (X_1, X_2, \ldots, X_p)$ is given by:

$$\Pr(Y = 1|\mathbf{x}) = \pi(\mathbf{x}) = \frac{\exp(\beta_0 + \mathbf{x}'\beta)}{1 + \exp(\beta_0 + \mathbf{x}'\beta)}.$$

where $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown coefficients.

Due to the great number of explanatory variables potentially involved in the estimation process, a penalized estimation technique has been employed to improve the interpretability of the final estimated model.

Given the sample $\{(\mathbf{x}_i', y_i), i = 1, 2, \ldots, n\}$, the log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \log(\pi(\mathbf{x}_i)) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i))\}$$

and the elastic net estimate is given by

Table 1 Table of covariates

Class	Variable	Short Description	Type
High school	Type of diploma	High school type (Classical studies, Technical, Scientific, ...)	Nominal
	Final exam mark	Total marks (from 60 to 100)	Integer
	Geographic area where the diploma has been obtained	Province and geographic area (the same province where the University is located, North, South or Center of Italy, abroad ...)	Nominal
Bachelor Degree	Course of study	BA, E, SAO	Nominal
	Enrollment age	University enrollment after the diploma: the next year of after one year, two or more years later	Nominal
	Enrollment motivation	Reasons behind the course of study choice	Nominal
	Graduation years	Number of years from the enrollment to the graduation	Integer
	Residence	Place of residence (in the same province as the University, different province but in the same region, different region, abroad)	Nominal
	Final mark	First degree final mark (from 66 to 110 cum laude; 110+laude=113)	Integer
	International experience	Participation to international exchange projects	Binary
	Satisfaction	Satisfaction of the University experience	Binary
	Organization	Satisfaction of the teaching and administrative organization (exams and lectures time table, information...)	Binary
	Teachers relationship	Satisfaction of the relationship with teachers	Binary
	Exams	Correspondence between the exam marks and the knowledge	Binary
	Library	Evaluation of the library services	Binary
	Equipments	Evaluation of the equipments used during the course of study (i.e. laboratories)	Binary
	Master wish	University where the student wishes to enroll for the master degree	Nominal
	Back to the start	The student is asked if, going back to its enrollment, he would have chosen the same University	Nominal
Socio-demo	Gender	Gender of the graduated student	Nominal
	Parents title	Title of study of the student parents	Ordinal
	Father title	Title of study of the student father	Ordinal
	Mother title	Title of study of the student mother	Ordinal
	Social status	Social status as coded by the AlmaLaurea survey	Nominal
Job	Job position	Current job position	Nominal
	Job province	The student is asked if he is willing to work in the same province where he resides	Binary
	Job Region	The student is asked if he is willing to work in the same region where he resides	Binary
	Job Center	The student is asked if he is willing to work in the Center of Italy	Binary
	Job South	The student is asked if he is willing to work in the South of Italy	Binary
	Job Europe	The student is asked if he is willing to work in Europe	Binary
	Job no-Europe	The student is asked if he is willing to work in a non-European country	Binary

$$\hat{\beta} = \arg\max_{\beta} \left\{ \ell(\beta) + \lambda \left[(1-\alpha) \sum_{j=1}^{p} \beta_j^2/2 + \alpha \sum_{j=1}^{p} |\beta_j| \right] \right\}.$$

The previous penalized estimate includes the LASSO as a special case when $\alpha = 1$. However, it can overcome some its well-known limitations. By combining L_1 and L_2 penalty terms, it can generates sparse models (L_1 term) and encourages grouping of correlated variables while stabilizing the L_1 regularization path (L_2 term). All computations have been made in R, by using the packages `glmnet` for the elastic net estimation ([5]) and `glmUtils` for the tuning of the parameters λ and α.

4 Results and conclusions

Table 2 shows the coefficients' estimates and the odds ratios for the relevant variables, that are selected by the elastic net for two values of $\lambda = (\lambda_{min}, \lambda_{1se})$ and for the optimal value of α. We can note that only 12 covariates have been chosen as relevant by the elastic net algorithm.

The variables with positive effect on the probability of continuing the study at the University of Salerno are the courses of study, the type of diploma, students willing to work in the same province and same region where they live. This means that the odds ratio is greater than 1, and therefore the probability of studying at the University of Salerno is greater than that of going to another university. By taking into account the reference group, we can note that students in Economics and also in Administration and Organization have a higher probability of studying at the University of Salerno with respect to the reference group, given by the students in Business Administration. This probability is much larger for Economics and for $\lambda = \lambda_{1se}$. Furthermore, the category Administration and Organization is not relevant at λ_{min}.

Then, students with technical and professional diploma have a churn risk lower than that for students with diploma in classical studies. Students who are willing to work in the same province and region where they live are more probably to continue their study at the University of Salerno. The two variables are relevant for both values of λ.

The variables with negative effects are social status, enrollment age, residence, international experience, satisfaction of the course of study and of the relationship with professors, students' choice if they could go back to their enrollment, students willing to work in no-European countries. This means that the risk of churn is higher for students i) with high social status; ii) who enrolled in the same year of diploma or with a year of delay; iii) who live in the same region where they study; iv) who had international experience during their study; v) who were not satisfied with the study experience and the relationships with professors; vi) who were willing to work in no-European countries; and vii) who would like to repeat their study experience by enrolling in the same course, both in the same university and in different universities.

The proposed model strategy allows to draw students' profiles with a high churn risk by evaluating the probability of continuing their study at the University of

Salerno on the basis of a relevant set of features collected. The characteristics that mainly affect the risk of abandonment are related to students' background (individual and familiar) and satisfaction of their study experience. As policy implications, only few factors, mostly connected to general students' experience, are under the control of university's governance to reduce the churn risk.

The current research could be improved by considering all departments (and consequently all courses of study) in the same university. Furthermore, the methodology described in this paper is able also to derive a measure of churn risk for master students.

Table 2 The coefficients' estimates and odds ratio for the significant variables selected by the elastic net, for two values of λ. The symbol '$-$' means that the variable is not significant for that value of λ.

	$\hat{\beta}$		odds	
	λ_{min}	λ_{1se}	λ_{min}	λ_{1se}
Course of study : Economics	0.13	0.27	1.14	1.30
Course of study : Administration and Organization	-	0.11	-	1.12
Social Status : High	-0.21	-0.33	0.81	0.72
High school diploma : Technical-Professional	-	0.07	-	1.07
Enrollment age : the same year of diploma	-0.05	-0.33	0.95	0.72
Residence : Same region of the study	-0.13	-0.29	0.88	0.75
International experience : Yes	-0.33	-0.49	0.72	0.61
Satisfaction : No Satisfied	-0.18	-0.27	0.84	0.76
Back to the start : Same course, different University	-0.79	-0.89	0.45	0.41
Back to the start : Same course and University	-	-0.10	-	0.90
Teachers relationship : No Satisfied	-0.17	-0.22	0.84	0.80
Job province : Yes	0.17	0.23	1.19	1.26
Job region : Yes	0.20	0.27	1.22	1.31
Job no-Europe : Yes	-0.11	-0.19	0.90	0.82

Acknowledgements The authors wish to thank the University of Salerno and its Statistical Office for providing the data and allowing the use of the AlmaLaurea datasets.

References

1. Bacci, S., Bertaccini, B.: Assessment of the University Reputation Through the Analysis of the Student Mobility. Soc. Indic. Res. **156**, 363–388 (2021)
2. Enea, M., Attanasio, M.: La mobilità degli studenti universitari nell'ultimo decennio in Italia. In: De Santis G., Pirani E., Porcu M. (eds) Rapporto sulla popolazione: l'istruzione in Italia, pp. 43–58. Il Mulino, Bologna (2019)
3. Figini, F., De Quarti, E., Giudici, P.: Churn Risk Mitigation Models for Student's Behavior. EJASA, Electron. J. App. Stat. Anal. **2**, 37–57 (2009)
4. Giudici, P., De Quarti, E.: Statistical Models to Predict Academic Churn Risk. In: Fichet, B., Piccolo, D., Verde, R., Vichi, M. (eds.) Classification and Multivariate Analysis for Complex Data Structures, pp. 41-49. Springer, Heidelberg (2011)
5. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. **33**(1), 1–22 (2010)
6. Tosi, F., Impicciatore, R., Rettaroli, R.: Individual skills and student mobility in Italy: a regional perspective. Reg. Stud. **53**, 1099–1111 (2019)

Discovering archetypal universities in higher education mobility flows in Italy

Analisi delle università archetipiche nei flussi di mobilità degli studenti universitari in Italia

Ilaria Primerano, Francesco Santelli and Cristian Usala

Abstract The aim of this contribution is to identify the archetypal universities in the Italian students' mobility network in terms of their attitude in attracting students. We define a set of networks according to the disciplinary groups by relying upon administrative data regarding students' mobility between bachelor's and master's degrees. For each disciplinary group, a network has been defined by considering the universities as nodes and the flows of students moving between nodes as links. Then, in each network, the set of archetypal universities is based on several network centrality indexes. Finally, these archetypes are used as benchmarks to identify the main determinants of universities' performances.

Abstract *L'obiettivo dell'analisi è quello di identificare le università archetipiche all'interno della rete di mobilità degli studenti universitari sulla base della loro capacità di attrarre o meno studenti. Sulla base dei dati amministrativi riguardanti le scelte di mobilità degli studenti tra la triennale e la magistrale, le singole reti sono state definite considerando, per ogni gruppo disciplinare, le università come nodi, e il flusso di studenti tra i nodi come legami. Per ogni gruppo disciplinare, è stato identificato il set di università archetipiche a partire dagli indici di centralità delle reti. Infine, questi archetipi sono stati utilizzati come riferimento per analizzare le determinanti delle performance delle università.*

Key words: University mobility, Archetypal Analysis, Network centrality indexes, Multivariate data

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

Francesco Santelli
Department of Political Sciences, University of Naples Federico II, Italy e-mail: francesco.santelli@unina.it

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

1 Introduction

In the last decades, the Italian student migration propensity has increased, highlighting the main flows from South to Northern Italy [1] [2]. The study of mobility flows can help measure the attractiveness of Italian universities. Different research approaches have been developed to study the determinants of students' mobility by means of different statistical methods, form longitudinal analysis [3] to Network Analysis considering geographical macro-area aggregations [4] [5], and different fields of study [6]. Most of the works refer to the first-level mobility, i.e. from high school diploma to bachelor degree, and only few recent contributions have dealt with the students' migration in the transition from Bachelor to Master degree programmes (i.e. second-level mobility), by focusing on Southern Italian students [7]. In studying second level mobility flows, a multilevel multinomial logit model has been adopted to assess the effects of university centrality role in the network (i.e. in terms of Hub and Authorities) on students' choices [6], while a multiplex network approach has been used to highlight the presence of groups of universities that play a fundamental role in each layer through the detection of the core universities [8].

Moving from this framework, this contribution investigates the Italian second-level mobility network, by considering the flows traced by students who change university for their master's degree by accounting also for differences existing among disciplinary fields. Specifically, a student who decides to change university when enrolling at the master is here considered as a student in mobility. Students' flows are analyzed to assess the similarity among Italian universities by identifying subgroups of universities within each disciplinary field that share common behavior in terms of their attitude in attracting students. In this context, students' flows are read into the scope of network analysis and, starting from a set of network centrality measures, the Archetypal analysis is applied to identify groups of similar universities within each disciplinary field.

The contribution is structured as follows: Section 2 describes the methodological approach; Section 3 presents an overview on the dataset used and the main results.

2 Network definition, centrality measures and archetypes

Based on Social Network Analysis [9], we start visualizing and analyzing Italian students' flows to get descriptive insights into the different network structures defined for each disciplinary field. We consider as nodes of source the universities where students achieved their bachelor's degree, and as nodes of destinations the universities where students enrolled for their master's degree. The flows of students moving in the second-level mobility network define the links connecting the Italian universities.

Specifically, starting from the ISCED-F 2013 classification, we define a set of one-mode, weighted, and directed networks [9]. Formally, each of these networks can be described as a graph $\mathcal{G}_i(V,L,W)$, where V is the set of units, L is the set of

directed links, and W is the set of weights. Let $\mathbf{A_i}$ be the corresponding adjacency matrices, with elements a_{ij} holding the presence of one or more links between node of origin v_i and node of destination v_j, otherwise $a_{ij} = 0$.

To gain insights on the characteristics of these networks, for each disciplinary field we have computed several network centrality indexes to measure universities' attractiveness and to determine whether the observed universities are top receivers of incoming students or, on the opposite, top senders of outgoing students. To classify the universities, we have defined a matrix holding by rows the Italian Universities and by columns the set of centrality indexes computed. This matrix is then used to perform an Archetypal analysis [10].

Archetypal analysis is a method of unsupervised learning that aims to represent each object in a dataset as a mixture of *individuals of pure type*, known as archetypes.Formally, archetypes are defined as m points $\{a_j\}$, $j = 1,...,m$ contained in archetype matrix $\mathbf{A}$ that are in the euclidean space that satisfy $x_i' = \alpha_i'\mathbf{A}$. The computation of the archetypes is a non-linear least squares problem, which is solved using an alternating minimizing algorithm. The key idea is that, given a $\mathbf{X}_{(n\times p)}$ data matrix, with n individuals and p variables, each archetype is computed as a linear combination of the original data based on the following constraints: $\beta_{ji} \geq 0$; $\sum_j \beta_{ji} = 1$; $\alpha_{ij} \geq 0$; $\sum_{i=1}^{m} \alpha_{ij} = 1$. The optimal α_{ij} are found by minimizing the following:

$$RSS = \sum_i \left\| x_i - \sum_{j=1}^{m} \alpha_{ij} A \right\|^2 \tag{1}$$

where α_{ij} are the coefficients of the archetypes while β_{ji} are the coefficients of the data set. Archetypes are useful in unsupervised learning also due to their location properties. Given a Convex Hull (CH) of original data points, if $k = 1$, only one archetype is identified, and the sample mean is the solution to minimize RSS; if $1 < k < n$, all the a_j vectors of archetypes lie on the boundary of CH to minimize RSS; if $k = n$ the number of archetypes is equal to n the $RSS = 0$ [10]. Using both sets of coefficients α_{ij} and β_{ji}, the RSS in (1) can be written using matrices notation, and thus matrix Γ including all the α's, matrix B of the coefficients β's and the starting matrix X of the data, obtaining in Frobenius norm:

$$\min_{\boldsymbol{\Gamma}_k, \mathbf{B}_k} RSS_k = \min_{\boldsymbol{\Gamma}_k, \mathbf{B}_k} \left\| \mathbf{X} - \boldsymbol{\Gamma}_k \mathbf{B}_k^T \mathbf{X} \right\|_F \tag{2}$$

In the space spanned by archetypes by exploiting the property of the compositional space defined by the Aitchinson distance [11], a k-means cluster analysis is performed to find consistent group of universities with respect to the archetypes.

3 Data description and main findings

We have collected data on students' mobility choices from the micro-data database MOBYSU.IT [12] that includes information on students university careers.[1] We consider the population of Italian students enrolled in a bachelor program in an Italian university between a.y. 2011-12 and a.y. 2016-17 that have enrolled in a master's degree program between the 2014 and 2019. We retain in our data only students that have graduated in this time frame (639,505 students, the 56.6% of the population), and that have enrolled in a master's degree program. Therefore, our data includes 400,049 students grouped in 92 universities (of which 10 are e-learning institutions). Moreover, to define the networks, we classify the students according to their disciplinary group. In particular, the set of available degree programs are classified into 10 disciplinary groups according to the ISCED-F 2013 classification [13]. Data are described in Table 1.

Table 1 Descriptive statistics on master's students mobility choices. Students are classified as *in mobility* if they have changed their university in the transition from bachelor's to master's degrees

	Master students		In mobility students	
ISCED - F 2013	Total N	In mobility N (%)	Same field (%)	Different field (%)
All fields	400,049	110,442 (27.61)	85.30	14.70
By disciplinary field in origin university:				
N1 - Agriculture, forestry, fisheries and veterinary	12,385	3,132 (25.29)	82.25	17.75
N2 - Arts and humanities	71,807	22,649 (31.54)	82.59	17.41
N3 - Business, administration and law	62,539	16,896 (27.02)	85.36	14.64
N4 - Education	13,508	3,569 (26.42)	92.23	7.77
N5 - Engineering, manufacturing and construction	86,518	14,347 (16.58)	98.62	1.38
N6 - Health and welfare	12,390	4,719 (38.09)	76.52	23.48
N7 - Information and Communication Technologies (ICTs)	4,809	938 (19.51)	85.15	14.85
N8 - Natural sciences, mathematics and statistics	46,023	12,743 (27.69)	95.68	4.32
N9 - Services	14,149	4,720 (33.36)	36.71	63.29
N10 - Social sciences, journalism and information	75,921	26,729 (35.21)	76.12	23.88

In the following, to show the results of our analytical procedure, we focused on the Education Field. The results of archetypal analysis performed on the network measures show that the best solution found by minimizing the RSS is obtained with 6 archetypes. In this space, the clustering algorithm identifies 12 groups of universities. The archetypes description with respect to original indexes (figure 1) and the clustering analysis (figure 2) show that the 6^{th} archetype has a peculiar trait with very high values for both attractiveness and exporting indexes (Roma Tre University is very close to this archetype), while the 5^{th} archetype represents an extreme type of university characterized by few flows (some Telematic universities are very

[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.

close to this archetype). Archetype 4^{th} is characterized by a greater attitude to export rather than import students (some Southern Universities), while the 3^{th} archetype, which is characterized by a very high value in both in–strenght and in–closeness network measures, fully describes University Pegaso and it is very close to Milano Bicocca, Bologna and Milano Cattolica.
The archetypal analysis on the network indexes highlights the peculiar behavior of the on–line universities in the mobility flows, and simultaneously a strong geographical component clearly emerges in most of the identified clusters.

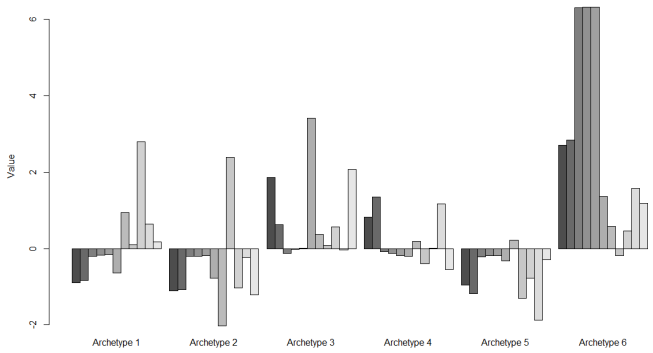


Fig. 1 Barplots for each archetype with respect to original variables (as archetypal coefficients). Columns are, from left to right: In–strength, Out–strength, Hub, Authority, Eigen–centrality, Page–rank, In–closeness, Out–closeness, Vertex–betweenness, N. of Recursive Deps, N. of reverse Recursive Deps

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. & Priulla, A.: Quali atenei scelgono i diplomati del Mezzogiorno d’Italia?. Neodemos, ISSN: 2421-3209, (2019)
2. D’Agostino, A., Ghellini, G., & Longobardi, S.: Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. Electron. J. App. Stat. Anal. **12(4)**, 826–845, (2019)
3. Attanasio, M., Enea, M., & Albano, A.: Dalla triennale alla magistrale: continua la ‘fuga dei cervelli’ dal Mezzogiorno d’Italia. Neodemos, ISSN: 2421-3209, (2019)
4. Columbu, S. & Primerano, I.: A multilevel analysis of university attractiveness in the network flows from bachelor to master’s degree. In A. Pollice, N. Salvati, and F. Schirippa Spagnolo, editors, Book of short Papers SIS 2020, pages 480–485, (2020)

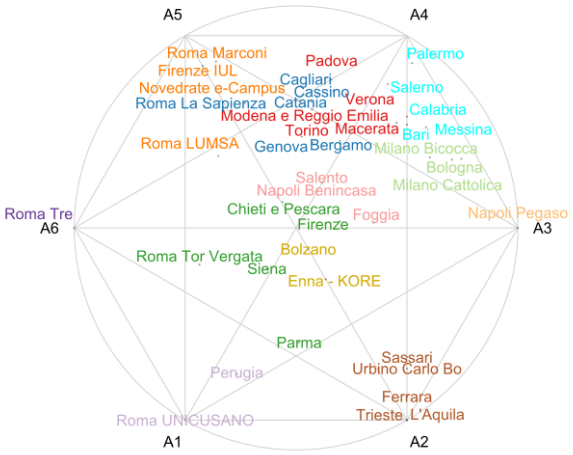


Fig. 2 Simplex plot of the space spanned by 6 archetypes. Universities are row–points. Clustering is performed to obtain 12 groups

5. 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)
6. Columbu, S., Porcu, M., Primerano, I., Sulis, I., & Vitale, M.P.: Geography of Italian student mobility: A network analysis approach. Socio-Econ. Plan. Sci. **73**, 100918 (2021)
7. Enea, M.: From South to North? Mobility of southern italian students at the transition from the first to the second level university degree. In C. Perna, M. Pratesi, and A. Ruiz-Gazen, editors, Studies in Theoretical and Applied Statistics, 239–249, (2018)
8. Primerano, I., Santelli, F., Usala, C.: A multiplex approach to study Italian Students' Mobility. In C. Perna, N. Salvati, and Schirripa Spagnolo F. (a cura di), Book of Short Papers SIS 2021, 473-478, (2021)
9. Wasserman, S., & Faust, K.: Social network analysis: Methods and applications, Cambridge University Press, (1994)
10. Cutler, A. & Breiman, L.: Archetypal analysis. *Technometrics* **36**, 338-347 (1994)
11. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. & Pawlowsky-Glahn, V.: Logratio analysis and compositional distance. *Mathematical Geology* **32**, 271-275, (2000)
12. Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenza, Cattolica and Napoli Federico II, Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA.
13. UNESCO Institute for Statistics. ISCED Fields of Education and International Standard Classification of Education 2011, Montrèal, (2014)

Measuring quality of students' careers in Higher Education: a systematic literature review

La misurazione della qualità delle carriere universitarie degli studenti: una revisione sistematica della letteratura

Clelia Cascella[1] and Giancarlo Ragozini[2]

Abstract Quality of student careers in Higher Education is a complex, multidimensional concept. Measuring it thus calls for the employment of an array of indicators and statistical techniques. The current paper is two-fold as it aims to (i) provide an updated definition of 'quality of student careers' (QoSC); and (ii) reconnoitre both the indicators used to operationalise such a concept and the statistical methods employed to assess student careers in Higher Education. To this end, we employed the PRISMA model to carry out a systematic literature review of both scientific and grey literature, written in English. The definition of QoSC was critically discussed along with the suitability of both the indicators and the statistical methods employed to measure QoSC, in contemporary societies (during/after Covid-19).

Abstract *La qualità delle carriere degli studenti universitari è un concetto complesso e multidimensionale. Misurarlo richiede l'impiego di una vasta gamma di indicatori e tecniche statistiche. Con questo articolo, presentiamo (i) una definizione di 'qualità delle carriere', e (ii) una rassegna degli indicatori e dei metodi utilizzati per operazionalizzare e misurare tale concetto. A questo scopo, abbiamo condotto una revisione sistematica della letteratura in lingua inglese, secondo il modello PRISMA, senza limiti di tempo. La definizione di qualità delle carriere è stata discussa così come l'adeguatezza degli indicatori e dei metodi impiegati per misurare tale concetto nelle società contemporanee (during/after Covid-19).*

Key words: quality, student careers, Higher Education

[1] Clelia Cascella, Ricercatore Confermato in Statistica Sociale presso l'Istituto Nazionale per la Valutazione del Sistema di Istruzione e Formazione (INVALSI), clelia.cascella@invalsi.it

[2] Giancarlo Ragozini, Professore di Statistica Sociale presso l'Università degli Studi di Napoli Federico II, giragoz@unina.it

1. Introduction

Scholars' and policymakers' interest in 'quality' in Higher Education (HE) has been broadly debated. It has been and still is at the top of the policy agenda worldwide (Bloch, Degn, Nygaard, and Haase, 2021). Nonetheless, relatively little research has been carried out so far about the concept of 'quality of student careers' (QoSC) in Higher Education and it has been mainly considered just as an indicator of the broader concept of 'quality of HE' institutions' (HEi), rather than a stand-alone subject.

Nonetheless, even though it is clear that university services are ancillary to quality of student careers in HE, previous studies have suggested that there are other factors that can affect student careers. The current paper aims to reconnoitre such factors and to provide an overview of the methodological approaches and analytical framework within which these factors have been operationalised and measured so far.

To this end, we carried out a systematic literature review by employing the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page, et al. 2021).

Results from our literature search were primarily intended to identify the current branches of research about and around the concept of 'quality of student careers' in Higher Education, in the attempt to provide an updated synthesis of the existing studies related to such a concept and to provide an updated definition of QoSC, able to mirror its complexity and multidimensionality in contemporary societies. Then, since the multidimensionality of QoSC calls for the employment of different methodological choices and analytical frameworks – including for example testing, measurement and assessment along with multivariate statistical analysis, indices construction, and so on – the current paper also aims to provide an overview of the statistical methods used to measure it. Finally, we critically discussed the definition of QoSC (based on the existing studies) and suggested some possible future research guidelines along with a critically review of the existing statistical methods employed to measure the 'quality of student careers' before and after Covid-19 outbreak.

2. Methods

A systematic literature review was conducted in 2021. The search terms "quality" AND "student" OR "university" AND "career(s)" AND "Higher Education (HE)" OR "Higher Education (HEi)" OR "Higher Education Institution (HEi)" were used to search publications written in English, available in ERIC - one of the largest online library of education research and information, sponsored by the Institute of Education Sciences (IES) of the U.S. Department of Education -, with no time constraints.

Article title, abstract and keywords were searched. Both academic (i.e. journal papers, books, and chapters) and grey literature (i.e., scientific reports) have been included.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), consisting of three steps ((i) identification, (ii) screening, and (iii) inclusion), was employed.

Identification. In Eric, we found 132,694 records published from 2002 to date. Among them, 4,997 records were published in 2021, 13,569 in 2020 and 39,815 in the last five years (i.e., from 2017 to date). In the present study, we focused on the records published in the last 5 years to provide an updated definition of QoSC.

Screening. To screen the publications (from the first step), we looked at both the title and the abstract to focus on the publications about 'quality of student careers' rather than about 'quality in Higher Education' or 'quality of HEi' (that include, for example, services provided by the university). After the screening, just 43 publications were put forward into the next step for further consideration.

Inclusion. In addition to the title and the abstract, we analysed the methodological section along with discussion and conclusion to identify the contribution to knowledge of each publication. Publications proposing a methodologically sounded reflection on the measurement of 'quality of student careers in Higher Education' were included. All the other publications were catalogued but not included in the current paper. The remaining 15 publications were moved forward into the review process.

Finally, the literature search carried out in ERIC was complemented via a snowballing search based on (i) the studies cited in the selected 15 publications, and (ii) by using the 'CITATION' tool available in Google.Scholar.

3. Results

When we think about QoSC, the first keyword that appears in our mind is likely to be 'achievement', measured for example via the average of the marks and/or the final grade, via the regularity throughout the academic pathway (e.g., the number of exams passed each year, the number of six- or less - years graduations), and so on. These factors have been used in previous studies to measure the quality of individual student career but have been – also and maybe foremost – taken as an indicator of the quality of (the services provided by) HEi, such as the number of six- (or less) years graduations, and, thus, as an indirect indicator of their attractiveness and competitiveness on the market.

Nonetheless, recent studies have focused on further aspects related to QoSC, such as the dropout rates (especially between the first and the second year, i.e., when the great majority of dropouts or decisions to transfer to another course of study occurs - Tinto 1975; Johnson 1997; Paura and Arhipova 2014), taken as an indicator of insufficient prospective student orientation (e.g., Perchinunno, Bilancia, Vitale, 2021), and thus as a predictor of students' success.

Other scholars focused on "long-distance outcome" and thus, for example, on the probability to find a (secure) job within 5 years after graduation and/or on the relationship between holding a degree and social mobility (the better the job and/or the wider the social mobility after graduation, the better the quality of student careers - e.g., Desai, 2019).

All the indicators mentioned above relate somehow to the concept of quality. The measurement of quality, by definition, implies a comparison between observed values and some standards. Such a perspective calls for a focus on measurable, comparable

outcomes (for example, the number of exams given by each student compared with the number of expected exams in the reference year). Nonetheless, there are further aspects related to the "quality of student careers" that involve students' experience in HE, such as students' perceptions about the quality of teaching (e.g., quality and quantity of feedback received by Lecturers and Teaching Assistants, quality of the teaching materials, office hours, and so on) and/or about the quality of services provided by the university (such as the library, laptop loans, support for housing, mental health, and so on). All these aspects do not directly relate to students' achievement but are ancillary to it (as the better the perceived quality, the better the students experience and, potentially, their achievements). Moreover, students' perceptions represent another, intangible aspect of student careers in HE thus expanding the concept of quality (by including other than measurable students' achievements). At our best knowledge, there are no studies including all the factors mentioned above. Nonetheless, the combined reading of existing studies seem to suggest that all these factors can be considered as interrelated sub-dimensions of QoSC: by a side, students' perceptions (the blue cells in the graph below) are ancillary to students' achievements (the orange cells); by the other side, students' achievement affects students' perceptions thus originating a multidimensional concept consisting of highly interrelated sub-dimensions (Figure 1).

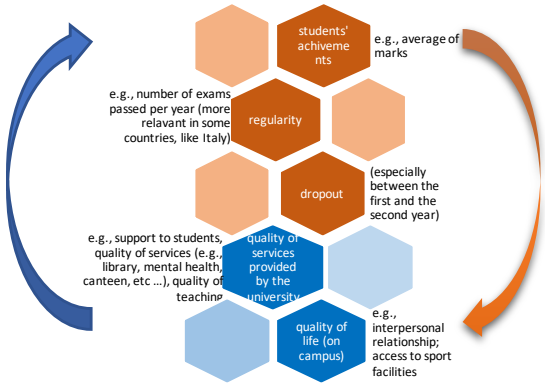


Figure 1. Dimensions of 'quality of student careers'

Compared to the studies focusing on students' achievements, the number of studies focusing on students' perceptions (e.g. Nabaho, Aguti, and Oonyu, 2019) is relatively scant, with just some exceptions especially in relation to some peculiar groups of students, such as minorities and, more frequently, in relation to students with disabilities (e.g. TEQSA, 2020)

Table 1 reports on a description of the aspects of QoSC included in the current literature along with a description of both the variables used to operationalise QoSC and the statistical methods employed to measure it. Via our literature search, we identified three main thematic groups, related to (i) students' achievement (including performance and long-distance outcomes), (ii) dropout rates, and (iii) students' perceptions/satisfaction (in relation to the services provided by the university and to life on campus). The studies focusing on students' achievements (group 1) are predominant and employ a wide array of methodological methods that include (i) testing, measurement and assessment, mainly within the framework of the IRT

modelling to measure students' learning outcomes, or (ii) (logistic and/or multilevel) regression to estimate, for example, the probability of getting a job after graduation - e.g. Desai, 2019 -, or to account for data hierarchy (e.g., students nested into faculties, universities, regions, countries). The study of dropouts (group 2) has mainly focused on two approaches: (i) the first is based on the observation of dropouts, at system/university level, carried out "to facilitate the identification of the most appropriate policy guidelines to reduce dropout rates in future cohorts" (Perchinunno et al., 2021); (ii) the second, generally known as 'Educational Data Mining' (EDM), has recently emerged and it is linked to the 'Churn analysis' that aims to predict the probability of dropout for each student on the basis of his/her characteristics (e.g., Ismail et al, 2015; Khodabandehlou and Zivari Rahman 2017). The Data Mining process, also known as 'Knowledge Discovery in Databases' (KDD), consists of the automatic discovery through appropriate algorithms of new and potentially useful information hidden within large amounts of data. It is thus used to discover regularities and new information within databases from contexts related to education, aimed at better understanding the individual students and the environments within which this instruction is provided, as well as their relation to the expected performance and objectives (e.g., Baker and Yacef 2009; Miguéis et al. 2018). The nature of these methods, and their relationship to classical statistical inference, is discussed by Perchinunno et al (2021). We grouped the other studies – focusing on a broader definition of quality in HE that encompasses 'quality teaching and learning', 'quality of facilities, services, and resources' (e.g., Iroegbu and Etudor-Eyo, 2020) as perceived by students (e.g., Calma and Dickson-Deane, 2020) – into the third group.

Table 1: Quality of student careers (QoSC) in previous studies

QoSC (sub-dimensions)	***Examples of sub-dimension definition and operationalisation***	***Method(s)***
Group 1: Students' achievements	Students' achievements are compared with the expected students' outcomes for example in terms of regularity throughout the academic pathway (e.g., number of exams passed per year), average of marks, probability of finding a job after graduation.	(Logistic) (Multilevel) Regression / IRT modelling
Group 2: Dropout rate	Percentage change between the number of students enrolled in two consecutive years (especially between the first and second year). Focus on the individual-level characteristics/decisions that can be used to increase the retention rate.	Knowledge Discovery in Databases (KDD) / Educational Data Mining (EDM) Churn analysis techniques (churn or attrition rate)
Group 3: Perceived quality	Students' perceptions about the possibility of implementing individual educational trajectories, migratory moods, and plans after graduation. Perceived quality of teaching Perceived quality of the services provided by university (including the library, support to students – e.g., laptop loans, mental support, housing, canteen, sport, and other facilities)	Factor analysis / (Self-administered) (Likert scale) questionnaire / Rasch or other IRT modelling / (composite) index

Source: our elaboration

4. Conclusion

QoSC is a complex and multidimensional concept. It includes a variety of aspects that are not just limited to the assessment of students' achievements (i.e., the focus of more than half of the studies carried out so far about the quality of student careers in Higher Education), but also includes dropout rate – conceived as an indicator of bad students' orientation –, or, more recently, students' perceptions/experience at the university.

The latter is relatively under-researched, but it may represent an interesting guideline for future research about the quality of student careers in HE, especially during / after Covid-19 outbreak. Even though the study of QoSC is clearly linked to the assessment of the quality of HEi (as QoSC can be taken as an indicator of the competitiveness of a university and thus, for example, as a measure of students' attractiveness), the study of QoSC cannot ignore the subjective components of quality as students' perceptions/experience has to be considered as a "primary component of quality" (e.g., Ruggeri, Warner, Bisoffi and Fontecedro, 2001).

In this perspective, QoSC should include both objective and subjective indicators, especially in contemporary societies, during and after Covid-19 outbreak that has established (at least a partial) transition from in-person to on-line or blended teaching protocols, especially in Higher Education.

In light of this, the methodological approaches and the analytical framework within which QoSC is investigated should not ignore the subjective components of QoSC and should move towards the development of an integrated approach that, at our best knowledge, has not yet been developed neither from a theoretical nor from a methodological perspective.

References

1. Aleshkovski, I.A., Gasparishvili, A.T., Krukhmaleva, O.V., & Onosov, A.A.: Students' Perceptions of Quality in Higher Education and Career Choices: A Case Study of the Russian Industrial Region. Eur. J. Cont. Edu **9**(4), 710-725 (2020).
2. Baker, R., & Yacef, K.: The state of educational data mining in 2009: A review and future visions. J. Edu Data Mining **1**(1), 3–17 (2009).
3. Bloch, C., Degn, L., Nygaard, S., & Haase, S.: Does quality work work? A systematic review of academic literature on quality initiatives in Higher Education. Ass. & Eval. HE **46**(5), 701-718 (2021).
4. Calma, A., & Dickson-Deane, C.: The student as customer and quality in Higher Education. Int. Jour. Edu. Man. **34**(8), 1221-1235 (2020).
5. Desai, A.: Refocusing Higher Education on Career Outcomes. Manhattan Institute for Policy Research (2019).
6. Ismail, M.R., Awang, M.K., Rahman, M.N.A., & Makhtar, M.: A multi-layer perceptron approach for customer churn prediction. Inter. J. Mult. Ubiq. Eng. **10**(7), 213–222 (2015).
7. Johnson, J.L.: Commuter college students: What factors determine who will persist or who will drop out? Coll. Stud. J. **31**(3), 323–332 (1997).
8. Khodabandehlou, S., & Zivari Rahman, M.: Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. J. Syst. Inf. Tech. **19**(1/2), 65–93 (2017).
9. Miguéis, V., Freitas, A., Garcia, P.J., & Silva, A.: Early segmentation of students according to their academic performance: A predictive modelling approach. Dec. Supp. Syst 115, 36–51 (2018).

10. Noaman, A.Y., Ragab, A.H.M., Madbouly, A.I., Khedra, A.M., & Fayoumi, A.G.: Higher Education quality assessment model: towards achieving educational quality standard. St. High. Edu. **42**(1), 23-46 (2017).
11. Nabaho, L., Aguti, J.N., & Oonyu, J.: Unravelling quality in Higher Education: what say the students? Africa Edu Rew **16**(5), 102-119 (2019).
12. Page M.J., McKenzie J.E., Bossuyt P.M., Boutron I., Hoffmann T.C., Mulrow C.D., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ **372**(71), 1-9 (2021).
13. Paura, L., & Arhipova, I.: Cause analysis of students' dropout rate in Higher Education study program. Procedia - Social and Behavioral Sciences **109**, 1282–1286 (2014).
14. Perchinunno, P., Bilancia, M., & Vitale, D.: A statistical analysis of factors affecting Higher Education dropouts. Soc. Ind. Res. **156**(2), 341-362 (2021).
15. Ruggeri, M., Warner, R., Bisoffi, G., & Fontecedro, L.: Subjective and objective dimensions of quality of life in psychiatric patients: A factor analytical approach: The South Verona Outcome Project 4. Brit. Jour. Psyc. **178**(3), 268-275 (2001).
16. Tertiary Education Quality and Standards Agency (TEQSA): Good practice note: Improving retention and completion of students in Australian Higher Education. Melbourne: TEQSA Ed. (2020).
17. Tinto, V.: Dropout from Higher Education: A theoretical synthesis of recent research. Rev. Edu. Res. **45**(1), 89–125 (1975).

Estimating the peers effect on students' university choices

Stima dell'effetto dei pari sulla scelta dell'università

Mariano Porcu, Isabella Sulis and Cristian Usala

Abstract This paper investigates the relationship between the probability of a student to enroll in a university away from home and the choices of her/his high school peers. Non local universities are defined accounting for the tertiary educational supply in students' local area and their subject of study, while the group of peers is identified as the set of enrolled students that have obtained their diploma in the same high-school, year and disciplinary field. For this aim, a two stages procedure has been adopted to disentangle the effect of individual and peers' demographic characteristics and educational backgrounds from the influence of peers' choices. The approach allows us to estimate the size of the peer effect as well as how this effect changes according to students' characteristics and the field of study.

Abstract *Questo lavoro analizza la relazione tra la probabilità degli studenti di iscriversi in una università distante dalla loro residenza e le scelte fatte dai propri pari. Le università sono definite come non locali tenendo in considerazione l'offerta di corsi universitari nell'area di residenza dello studente e il campo di studi scelto, mentre i pari sono idenficati come tutti gli studenti immatricolati che hanno ottenuto il loro diploma nella stessa scuola superiore, lo stesso anno e nello stesso indirizzo. Per la stima è stata utilizzata una procedura a due stadi che permette di separare l'effetto legato alle caratteristiche dei pari e della scuola frequentata da quello legato solamente alle scelte. L'approccio utilizzato permette la stima della dimensione dell'effetto e di valutare come lo stesso si modifica al variare delle caratteristiche degli studenti e del campo di studi scelto.*

Key words: higher education, mobility choices, peer effect, strategic interactions

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

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

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

1 Introduction

Studies carried out on students' university choices in the last decade indicate that students' decisions are mainly driven by factors which are not strictly connected to the academic prestige of the universities in terms of research activities and quality of the curricula supplied. The socioeconomic conditions of the places where the universities are located and their link with the local job market play a primary role in determining students' university choices. Family aspirations in terms of the future employment status of their children influence their decision to invest in tertiary education institutions which can ensure better employment opportunities, promoting the choice of those institutions which are located in the most dynamic job-markets, with well established networks of collaborations between universities and companies in the private and public sectors (see [1]). The national literature mainly focuses on quantifying the size and the direction of students' mobility flows between macro geographical areas, highlighting how this kind of mobility is strictly linked to the brain drain of human capital and play part in the consequent well-known migration chains which contribute to enlarge demographic and socioeconomic disparities between core and peripheral provinces/regions (see [2, 3]). Moreover, studies which focus on the analysis of the network configurations of such flows of students identify universities, regions and provinces which act as sender or receiving institutions or both in the students exchange network. In that way the main links between territories and institutions are highlighted as well as the main determinants of the observed divergences across heterogeneous entities (see [4, 5]).

Moving from this framework, this work aims to shed some lights on the influence that the secondary school environment has in addressing students' choices at the university. In particular, we analyze students mobility choices considering the choices of their peers by identifying as peers the set of freshmen that have obtained their diploma in the same high-school, year and disciplinary field. At this aim a two stages procedure has been adopted to disentangle the effect of individual and peers' demographic characteristics and educational background from the influence of peers' choices. In particular, this work take advantage of the methodology proposed by [6] to estimate the peers' effect in stock market analysts' recommendations in a framework of static games with strategic interactions. As far as the authors know this is the first work that applies this methodology to estimate the size of peers' influence on students' mobility choices.

2 Data

This analysis relies upon the micro-data on Italian university students extracted from the database MOBYSU.IT [7] which includes several information on univer-

sity students' careers, their individual characteristics and educational background.[1] In particular, we consider the population of students who enrolled in a bachelors' program in an Italian university between 2019 and 2020, and that have graduated at high school after the 2019. Since we are interested in modeling students' mobility choices, we do not consider students enrolled in programs accessible with a national entry test or in e-learning universities. Indeed, in the first case, students' choices are determined by their position in the national rank rather than by their preferences while, in the second, students do not have to move to reach their university. Moreover, in order to get a consistent definition of students' peers group, we do not consider the pupils that have attended their high school abroad or those for which we cannot identify the high school (e.g. students who did not report their high school curriculum). Therefore, starting from a population of 548,540 freshmen, we retain in our sample 407,701 students.

Students' university mobility choices are classified according to the tertiary education supply in their local area by accounting for the minimum travel time needed to reach the nearest university. In particular, we obtain the data regarding the minimum travel distance by car between any pair of Italian cities by combining the travel times obtained from ISTAT matrices and the data available on Google Maps. Then, for each student i, we computed the minimum distance from the nearest university d_i to distinguish between *local* and *non local* universities. Moreover, to consider that students may see as local also universities located close to the nearest university, we use a broader definition of students' local area by adding 30 minutes to each threshold d_i. Thus, from a student perspective, we define two categories of universities: *local* and *non local*. *Local* universities are those located less than $d_i + 30$ minutes of travel from students' city of residence, while *non local* universities are those located farther than $d_i + 30$ minutes of travel. This strategy allows us to classify students choices without relying simply on administrative borders or other arbitrary assumptions but accounting for the supply of universities in students' residence area. According to this definition, we have that the 31.9% of observed students have decided to enroll in a *non local* university, this share increases if we consider students living in the South (34.2%) and in Islands (37.3%), while is lower in the North (31.5%) and in the Centre (27.7%).

For each student we define her/his peers group as the set of pupils who have attended the same *class*. Each *class* includes all the freshman that have obtained their diploma in the same high-school, the same year and have attended the same curriculum. This definition of peers has two caveats. First, we do not observe the entire set of students' peers but only those that have enrolled in an Italian university. Therefore, we assume that students' mobility choices will depend only on those peers that have decided to enroll at university in Italy. However, since we are modeling the mobility choices of enrolled students rather than the enrollment choice this problem should not affect our results. The second problem is given by the fact that our

[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.

definition of *classes* is broader than the standard definition of classrooms. Indeed, we are not able to split the *classes* in sections and, therefore, our definition of peers include also the students that may have not attended their high-school in the same classroom intended in strict sense. Thus, we assume that the students are interacting not only with their classroom mates but also with all the other students that have attended the same curriculum in the same year at the same school. According to this definition, students are grouped in 35,575 *classes*, the median number of peers in a *class* is 24, the mean is 40, the 5th percentile is 5, the 95th percentile is 134.

3 Empirical framework

The utility of student i to enroll in a *non local* university ($Y_i = 1$), conditioned on her/his individual characteristics D_i, peers average characteristics G_{-i}, peers choices Y_j, and a private stochastic preference shock ε_i is modeled as:

$$U(Y_i = 1|Y_{-i}, D_i, G_{-i}) = \alpha + \delta \frac{\sum_{j \in -i} I(Y_j = 1)}{N_{-i}} + \Gamma' D_i + \Upsilon' G_{-i} + \varepsilon_i \qquad (1)$$

where $-i$ indicates the set of peers in the *class* attended by student i. The set D_i includes student's age, gender (1 = females) and no-native (1 = no-native) dummies, their school final grade, the distance needed to reach their school and the nearest university, their macro area of residence, lyceum and private school dummies, provincial unemployment and regional GDP. The set G_{-i} stands for the proportion of females and no-natives in the class computed without considering the student i.

Peers interactions are modeled following the approach developed in [6] using a static game with incomplete information in which peers takes their choices simultaneously. This assumption implies that students do not know peers' preference shocks and, therefore, they do not observe all the vector of peers' choices. Thus, students form expectations on their peers' choices based on their observed characteristics. In this framework, Eq. 1 can be rewritten to consider students' expected utility to choose a *non local* university as follows:

$$U_e(Y_i = 1|Y_{-i}, D_i, G_{-i}) = \alpha + \delta \frac{\sum_{j \in -i} \hat{\sigma}_j}{N_{-i}} + \Gamma' D_i + \Upsilon' G_{-i} + \varepsilon_i \qquad (2)$$

where $\hat{\sigma}_j$ indicates students' beliefs on peer j choices and the parameter δ informs on the effect of peers choices on student i utility. This model can be estimated using a two-stages procedure. In the first stage peers' expected choices $\hat{\sigma}_j$ are estimated through a sieve nonparametric regression by approximating students' choice probability with a set of flexible polynomials of student's own and peers' characteristics (see [8]). In our case, we have designed a set of Hermite polynomial up to the third degree of all the characteristics comprised in D_i and G_{-i} and their interactions. In the second stage, the Eq. 2 is estimated by using the average $\hat{\sigma}_j$ for each *class* as a measure of students' beliefs regarding peers' choices. Since the dependent variable

Table 1 Students' utility parameters

	Non Local vs Local				
	Parameter	Std Error		Parameter	Std Error
Peers effect, $\hat{\delta}$	4.982	(0.035)	Diploma grade	0.072	(0.004)
Age	–0.006	(0.004)	Lyceum	–0.010	(0.003)
Females	0.034	(0.007)	Private HighSchool	–0.077	(0.008)
No-native	0.029	(0.020)	Prov. Unemployment	–0.034	(0.004)
South	0.032	(0.007)	Reg. GDP	–0.053	(0.004)
Island	0.094	(0.007)	% females in class	–0.012	(0.008)
North	0.118	(0.004)	% no-native in class	–0.044	(0.027)
Min distance from Univ.	–0.080	(0.005)	Constant	–2.500	(0.000)
Distance from HighSchool	–0.025	(0.006)			
Observations	407701				
Pseudo R-squared	.0772817				

is binary, the second stage is estimated using a logit regression and the standard errors are recovered by using a bootstrap procedure with 200 replications. In this framework, the identification of peers effect relies on the assumption that some of the peers' characteristics affect peers' choices but not the utility of student i [9]. This exclusion restriction implies that students' utility do not vary with the age or the final grade of the other students that choose a *non local* university. Namely, the set of peers' characteristics G_{-i} do not contains all the variables contained in the set D_{-i}.

Table 1 reports the preliminary results of this two-stage procedure where the parameter that measure peers' influence on students' choices is indicated as $\hat{\delta}$. To ease the interpretation we have standardized all the variables comprised in D_i and G_{-i} in z-scores. As we can note, peers' choices have a strong and positive influence on students probability to choose a *non local* university. Indeed, for the average student, an increase in the average share of peers that choose a *non local* university of 5% (corresponding to 1 student over 20) is related to an increase in students' choice probability of 5.4%.

With respect to the other determinants, we can see that the probability to choose a *non local* university is positively related to the female indicator and the diploma grade, while is negatively related to the minimum distance from the nearest university and the distance traveled to reach the high school. This result indicates that, *ceteris paribus*, students that live in more isolated areas are more likely to choose a *non local* university to pursue their career. These results are even stronger if we consider the multiplicative effect of peers choices. Indeed, our results show that when the share of peers that choose a *non local* university increases, for reasons that may depend on students' diploma grade or distances, we have also an increase in individual choice probabilities of mobility. Therefore, this approach could shed some light also in understanding the mechanisms behind the chain migration phenomenons.

4 Conclusions

The methodology approach adopted allowed us to provide evidences on the role that peers' choices have in affecting students' decision process. Our results indicate that students' choice probability to choose a *non local* universities is positively affected by the share of peers that have made the same choice in their *class*.

Further development of the analysis includes the use of INVALSI micro-data to account for the effect of students socio-economic characteristics and the application of a multilevel approach that allows us to control also for the role played by schools. Moreover, this approach will be extended also to consider (i) the heterogeneity of peers' effect based on individual characteristics (e.g. diploma grade); (ii) divergences on the size of the peers effect across disciplinary fields; (iii) the heterogeneity in the tertiary education supply in students' local area in terms of degree availability. In this respect, our aim is to investigate the differences between *free* and *forced* mobility by considering in a separate category those students that have enrolled in a *non local* university which, however, is the nearest institution providing the chosen degree program.

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

References

1. Cattaneo, M., Horta, H., Malighetti, P., Meoli, M., and Paleari, S.: Effects of the financial crisis on university choice by gender. Higher Education **74(5)**, 775–798, (2017)
2. Ciriaci, D.: Does University Quality Influence the Interregional Mobility of Students and Graduates? The Case of Italy. Regional Studies **48(10)** 1592-1608, (2014)
3. Attanasio, M., Enea, M.: La mobilitá degli studenti universitari nell'ultimo decennio in Italia. In G. De Santis, and E. Pirani and M. Porcu, editors, Rapporto sulla popolazione. L'istruzione in Italia, Il Mulino, (2019)
4. Primerano, I., Santelli, F., Usala, C.: A multiplex approach to study Italian Students' Mobility. In C. Perna, N. Salvati, and Schirripa Spagnolo F. (a cura di), Book of Short Papers SIS 2021, 473-478, (2021)
5. Columbu, S., Porcu, M., Primerano, I., Sulis, I., and Vitale, M.P.: Geography of Italian student mobility: A network analysis approach. Socio-Econ. Plan. Sci. **73**, 100918, (2021)
6. Bajari, P., Hong, H., Krainer, J., & Nekipelov, D.: Estimating static models of strategic interactions. Journal of Business and Economic Statistics **28(4)**, 469–482,(2010)
7. Database MOBYSU.IT: Mobilità degli Studi Universitari in Italia. Research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II. Scientific Coordinator Massimo Attanasio (UNIPA). Data Source ANS-MUR/CINECA
8. Ai, C., and Chen, X.: Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica **71(6)**, 1795–1843, (2003)
9. Yakovlev, E.: Demand for alcohol consumption in Russia and its implication for mortality. American Economic Journal: Applied Economics **10(1)**, 106–149, (2018)

Session of solicited contributes SS4 – *Conformity assessment and quality predictions- itENBIS*
Organizer and Chair: Amalia Vanacore

Testing the predictive performance of multi-class classifiers

Valutazione delle prestazioni degli algoritmi di classificazione in problemi multi-classe

Amalia Vanacore, Maria Sole Pellegrino and Armando Ciardiello

Abstract The necessity of estimating classifiers predictive performance and testing whether they performs satisfactorily for the problem at hand are common needs in classification problems. This paper suggests to estimate classifier performance via Balanced Agreement Coefficients. The global hypothesis that all classifiers under study perform satisfactorily can be tested via max-*t* test. The suggested procedure is applied to test the predictive performance of three machine learning algorithms in a problem of ordinal classifications.

Key words: Multi-class classifications, Classifier predictive performance, Balanced Agreement Coefficient, Max-*t* test

1 Introduction

Two important aspects when comparing classifiers are estimating their predictive performance and testing whether all of them perform satisfactorily for the problem at hand.

Most of the available metrics for assessing classifier performance (see [9, 10] for an overview) are determined on the basis of confusion matrix, a cross table that records how cases are distributed over predicted (on columns) and actual (on rows)

Amalia Vanacore
Dept. of Industrial Engineering, University of Naples Federico II, p.le Tecchio 80, Naples (Italy)
e-mail: amalia.vanacore@unina.it

Maria Sole Pellegrino
Dept. of Industrial Engineering, University of Naples Federico II, p.le Tecchio 80, Naples (Italy)
e-mail: mariasole.pellegrino@unina.it

Armando Ciardiello
Dept. of Industrial Engineering, University of Naples Federico II, p.le Tecchio 80, Naples (Italy)
e-mail: armando.ciardiello@unina.it

classes in such a way that cells on the main diagonal count the correctly classified cases whereas off-diagonal cells the cases incorrectly assigned to the class.

The most widespread classifier performance measure derived from confusion matrix is the accuracy, which is the number of successful predictions relative to the total number of classifications. The main criticism raised against accuracy is that it may lead to erroneous conclusions with imbalanced data sets since it strongly depends on the performance over the majority classes. A more robust measure of classifier performance is the balanced accuracy, obtained by averaging the accuracy values estimated for each class. However, both accuracy and balanced accuracy do not compensate the non-zero probability that some classifications match only by chance.

An alternative measure of predictive performance able to compensate the effect of classifications matching by chance is Cohen's K coefficient [2]. It is a relative measure of agreement belonging to the family of κ-type coefficients introduced in social and behavioral sciences for measuring the degree of rater agreement and in the last decades also adopted as measure of classifier performance within the context of expert systems, machine learning and data mining communities [1, 4, 13]. Some authors regard Cohen's K as a more robust measure than accuracy because it estimates the probability of classifications matching by chance through marginal frequencies making the coefficient value decrease in the presence of imbalanced data sets. Actually, the chance-agreement term of Cohen's K coefficient produces a penalization rather than a direct and verifiable correction for imbalance and thus it is not clear how the coefficient balances the classifier performance over majority and minority classes. Moreover, it is worth to highlight that when the imbalance is asymmetrical between actual and predicted classes (i.e. the worst performance ever), the Cohen's K coefficient increases leading to a strongly misleading conclusion. For these reasons, Cohen's K should be avoided as measure of classifier predictive performance, especially with imbalanced data sets.

A robust measure of classifier predictive performance can be obtained by correcting the balanced accuracy with the proportion of classification matching by chance formulated as independent from the marginal frequencies. Thus, we propose the balanced version of the Agreement Coefficient developed by Gwet [5], hereafter denoted as Balanced AC_1. Indeed, the strategy for chance-correction in Gwet's AC, though defined in function of rater agreement problem, is recognized as relevant also for assessing classifier predictive performance due to its propensity to be weakly affected by marginal frequencies and thus suitable for managing all data sets [7]. Furthermore, the weighted variant of Balanced Agreement Coefficient, hereafter denoted as Balanced AC_2, allows to introduce a weighting scheme in order to account for the cost of misclassification severity, since in multi-class classification problems with an inherent order among classes some misclassifications are more expensive or harmful than others.

When investigating classifier predictive performance, different classifiers are considered, either on the same or different data sets [8, 11]. We focus here on the performance of several classifiers over the same data set and check, via max-t test [6], whether all classifiers under comparison perform satisfactorily with respect to a

desired prediction threshold value. The max-t test is a simultaneous inference procedure that considers the intersection of several individual null hypotheses and defines a common rejection region based on the joint distribution of test statistics.

The applicability and usefulness of the recommended Balanced AC_2 coefficient together with max-t test are illustrated through a real data set concerning ordinal multi-class classification problem. The predictive performance of three machine learning algorithms (i.e. Deep Neural Network, Random Forest, Extreme Gradient Boosting) assessed via Balanced AC_2 is compared against Accuracy, Balanced Accuracy and Balanced AC_1.

The paper is organized as follows: in Section 2 classifier performance measures and max-t test are introduced; both of them are applied to a real data set in Section 3 and finally conclusions are summarized in Section 4.

2 Assessing and testing classifiers performance

Let n be the number of cases classified on $k \geq 2$ classes, n_{ij} the number of cases with actual class i but predicted class j and w_{ij} the symmetrical agreeing weight for (i, j) cell of confusion matrix. The Accuracy and Balanced Accuracy are given by:

$$\text{Accuracy} = \frac{\sum_{i=1}^{k} n_{ii}}{n}; \qquad \text{Balanced Accuracy} = \frac{\sum_{i=1}^{k} n_{ii}/n_{i\cdot}}{k}. \tag{1}$$

The Balanced AC_2 coefficient, instead, is formulated as follows:

$$\text{Balanced } AC_2 = \frac{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}w_{ij}/n_{i\cdot}}{k} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(n_{ij}w_{ij}/n_{\cdot j})(1-n_{ij}w_{ij}/n_{\cdot j})(n_{ij}w_{ij}/n)}{k-1}}{1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(n_{ij}w_{ij}/n_{\cdot j})(1-n_{ij}w_{ij}/n_{\cdot j})(n_{ij}w_{ij}/n)}{k-1}} \tag{2}$$

where $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$. The unweighted coefficient Balanced AC_1 for multi-class classifications can be assessed via Eq. 2 with w_{ij} = 1 if $i = j$ and w_{ij} = 0 elsewhere.

The performance of different classifiers on a given data set can be checked against a desired threshold value via max-t test [6]. The tested system of hypotheses is:

$$H_0 = \bigcap_{m=1}^{M} H_{0,m} \quad \text{against} \quad H_1 = \bigcup_{m=1}^{M} H_{1,m} \tag{3}$$

where H_0 and H_1 are the global null and alternative hypothesis; $H_{0,m} : \theta_m \leq \theta_0$ and $H_{1,m} : \theta_m > \theta_0$ are the m^{th} individual null and alternative hypothesis; θ_m is the performance for the generic classifier $m = 1, \ldots, M$ and θ_0 is the common threshold performance value. Specifically, the max-t test is based on the approximate multivariate normal distribution of test statistics $t_m = (\hat{\theta}_m - \theta_0)/\hat{se}(\hat{\theta}_m)$, for $m = 1, \ldots, M$, and considers the intersection of several null hypotheses by defining a common re-

jection region accounting for the the correlation between t_m while controlling the Family Wise Error Rate (FWER).

Since the rejection of the global null hypothesis is favored by larger values of classifier performance $\hat{\boldsymbol{\theta}}$, H_0 is rejected only if the maximum among individual test statistics t_m equals at least a common critical value c_α calculated from the joint distribution of the test statistics by solving numerically the following equation:

$$P\left(\max_{m\in M} t_m \leq c_\alpha\right) \approx \int_{(-\infty, c_\alpha)} \phi_M\left(x, \hat{\mathbf{R}}\right) dx = 1-\alpha \tag{4}$$

where ϕ_M is the pdf of the M – dimensional normal distribution and $\hat{\mathbf{R}}$ is the estimated correlation matrix of $\hat{\boldsymbol{\theta}}$. When H_0 is rejected, the max-t test could provide information about which of the individual null hypotheses is significant, that is which classifiers have performance greater than the threshold value θ_0, by the construction of simultaneous confidence intervals with coverage probability $1-\alpha$:

$$\mathrm{CI}_{1-\alpha} = \left[\hat{\boldsymbol{\theta}} - \mathbf{c}_\alpha \cdot \hat{se}(\hat{\boldsymbol{\theta}}), \infty\right) \tag{5}$$

3 An illustrative example: wine quality prediction

A classification problem dealing with ordinal classifications has been solved by three different multi-class classifiers: Deep Neural Network (DNN), Random Forest (RF) and Extreme Gradient Boosting (XGB). The data set includes 1599 samples, divided into training and test data sets [3]. For each sample, 11 physicochemical variables (input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, ph, sulphates and alcohol) are provided together with wine quality score (sensory output variable) obtained as the median of at least 3 evaluations made by as many wine experts who graded the wine quality with an ordinal score ranging between 1 (very bad) and 10 (very excellent). The distribution of the wine quality score is very unbalanced and ranges between 3 and 8 (see Figure 1).

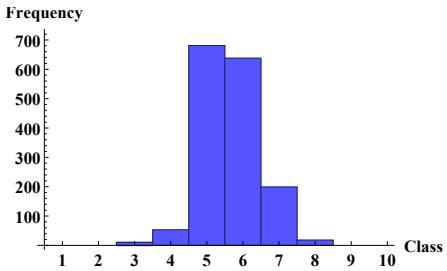


Fig. 1 Marginal distribution of wine quality scores

Due to the ordinal nature of quality evaluations, the classifier performance is assessed via the linearly weighted agreement coefficient and then compared against its unweighted variant.
The suggested method for performance estimation is based on Repeated Stratified Nested Cross-Validation (CV), a nesting of two repeated stratified V-fold CV loops: the inner loop responsible for model selection and the outer loop responsible for generalization performance estimation. A number of 10 repetitions are performed for both loops with V=3 and V=5 folds for inner and outer loop, respectively.
The estimates of the large-sample predictive performance given by the mean over the Nested CV performance values are reported (in bold) in Table 1 for each classifier together with the estimated interval of the large-sample predictive performance given by the interval between the minimum and maximum over the nested CV performance values (in square brackets).

Table 1 Estimates of classifier predictive performance

	Accuracy	**Balanced Accuracy**	**Balanced AC_1**	**Balanced AC_2**
DNN	**0.5583**	**0.2666**	**0.2466**	**0.7689**
	[0.470÷0.621]	[0.190÷0.307]	[0.171÷0.286]	[0.692÷0.817]
RF	**0.6818**	**0.3484**	**0.3295**	**0.7950**
	[0.375÷0.743]	[0.177÷0.410]	[0.162÷0.392]	[0.691÷0.834]
XGB	**0.6836**	**0.3444**	**0.3251**	**0.7898**
	[0.608÷0.722]	[0.278÷0.405]	[0.257÷0.388]	[0.753÷0.856]

The obtained results show the unsuitability of Accuracy for this data set because of its dependency on the performance over the majority classes which leads to an overestimate of classifier performance when compared against Balanced Accuracy. Moreover, it is worth to note the difference, although slight, between Balanced Accuracy and Balanced AC_1 due to the correction of the former by the proportion of classifications matching by chance. More interestingly, it is evident that the weighted coefficient Balanced AC_2 achieves greater values than the unweighted variant Balanced AC_1, highlighting the positive effect of weighting misclassifications according to their severity.

Focusing on Balanced AC_2, it is tested whether classifiers exhibit an almost perfect predictive performance via max-t test, fixing $\theta_0 = 0.8$. There is not evidence for rejecting the global null hypothesis with a FWER of 5% meaning that there is at least one classifier with a performance lower than the threshold value. In order to identify which $H_{0,m}$ can be rejected, the simultaneous confidence intervals have been built according to Eq. 5. The obtained results reveal that all individual null hypotheses cannot be rejected since the lower 95% confidence bounds are 0.7578, 0.7833 and 0.7784 for DNN, RF and XGB, respectively. The DNN is the classifier with the worst predictive performance; this result is not surprising since it is well known that neural networks need a greater number of observations to achieve better predictive performance.

4 Conclusions

As stated by *No free lunch theorem* [12] in machine learning, no one model works best for all possible situations; choosing performance measures blindly and applying them without any regard for their meaning and the conditions governing them is not a particularly interesting endeavor that can result in dangerously misleading conclusions.
This research study introduces Balanced Agreement Coefficients as novel measures of classifier predictive performance in multi-class (either nominal or ordinal) classification problems. The proposed measures account for chance classifications and are able to deal with imbalanced marginals, which are a rule in many real-word multi-class classification problems.

Future research work will be addressed to conduct simulation experiments in order to investigate the statistical behavior of the proposed classifier performance measures under different scenarios.

References

1. Ben-David, A.: Comparison of classification accuracy using cohen's weighted kappa. Expert Systems with Applications **34**(2), 825–832 (2008)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
3. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decision support systems **47**(4), 547–553 (2009)
4. Duro, D.C., Franklin, S.E., Dubé, M.G.: A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery. Remote sensing of environment **118**, 259–272 (2012)
5. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
6. Hothorn, T., Bretz, F., Westfall, P.: Simultaneous inference in general parametric models. Biometrical Journal: Journal of Mathematical Methods in Biosciences **50**(3), 346–363 (2008)
7. Labatut, V., Cherifi, H.: Evaluation of performance measures for classifiers comparison. arXiv preprint arXiv:1112.4133 (2011)
8. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808 (2018)
9. Sammut, C., Webb, G.I.: Encyclopedia of machine learning. Springer Science & Business Media (2011)
10. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information processing & management **45**(4), 427–437 (2009)

11. Stapor, K., Ksieniewicz, P., García, S., Woźniak, M.: How to design the fair experimental classifier evaluation. Applied Soft Computing **104**(107219), 1–12 (2021)
12. Wolpert, D.H.: The supervised learning no-free-lunch theorems. Soft computing and industry pp. 25–42 (2002)
13. Zhou, J., Li, E., Yang, S., Wang, M., Shi, X., Yao, S., Mitri, H.S.: Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories. Safety Science **118**, 505–518 (2019)

Optimal Subgrids from Spatial Monitoring Networks

Selezione di mappe ottime da una griglia di campionamento

Riccardo Borgoni, Andrea Gilardi and Diego Zappa

Abstract The monitoring of production processes on a planar surface typically involves sampling network to gather information about the status of the process. In order to save time and money, when the process goes into a stable status it might be appropriate to reduce the dimension of the sampling grid. In some cases, the allocation of a new network of smaller dimension is not free of constraints and it might be necessary the selection of a subgrid extracted from the original network. Discussion is focused on some recent methods used to achieve this aim. Possible extensions to consider jointly tabu search algorithm and co-kriging models is reported.

Abstract *I processi produttivi aventi come dominio una superficie richiedono, per il monitoraggio, l'impiego di mappe di campionamento. Al fine di contenere I costi della raccolta dei dati, quando il processo produttivo raggiunge una certa stabilità è spesso necessario ridurre la dimensione della mappa inizialmente calibrata. Questa operazione non sempre è priva di vincoli e potrebbe essere necessario dover scegliere un sottoinsieme di punti dalla mappa iniziale. La questione discussa in questo lavoro è quali criteri usare in questo contesto. Una proposta basata su una combinazione di algoritmi di "tabu search" uniti a modelli di spaziali di "co-kriging" viene proposta come possibile estensione ai metodi attualmente disponibili.*

[1] Borgoni Riccardo, University of Milano-Bicocca, Metodi Quantitativi e Strategia d'impresa; email: riccardo.borgoni@unimib.it
Gilardi Andrea, University of Milano-Bicocca, Metodi Quantitativi e Strategia d'impresa; email: andrea.gilardi@unimib.it
Zappa Diego, Università Cattolica del Sacro Cuore, Department of Statistical Sciences, Milano; email: diego.zappa@unicatt.it

Key words: Samplig grid, Process monitoring, Tabu search, Kriging

1 Intoduction

Productions are often monitored by gathering data form a monitoring network. Examples can be found in agriculture, textile, steel, microelectronics processes. Grids are usually allocated according to some optimal spatial criteria or ad hoc defined because of technological or physical constraints. When the production process is at an early stage, the size of the network can be moderately large but when the process becomes stable and volumes significantly increase, a reduction of the grid size is necessary. In case a new grid cannot be redesigned on the production surface it is necessary to select a subgrid from the initial one. That is the case when the aim is to compress sampling costs and to guarantee the link with the sampling locations used in the past. From a theoretical point of view, the best selection procedure should consider and compare all the possible subsets of a given size, test the predictive capability of the reduced grid and select the new optimal configuration. The drawback is that this is a formidable combinatorial problem even when the number of measurement locations is moderately high. For further details see [1][2].
The rest of the short paper contains a description of the problem, the solution so far adopted and possible extension to the tabù search algorithm in a multi objective context.

2 Subgrid selection

To focus on the problem, Fig.1 depicts real networks (amongst many others) adopted to monitor semiconductor processes.

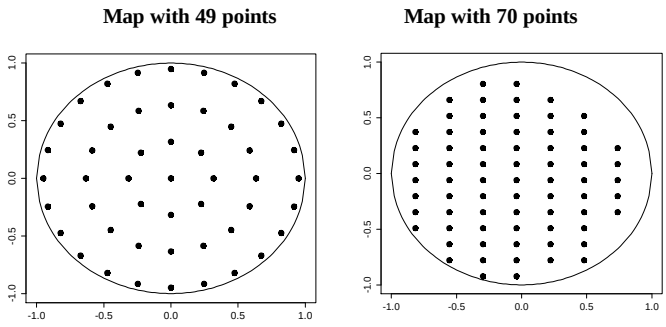


Figure 1: Monitoring grid used to monitor semiconductor processes

The matter is: is it possible to reduce the size of the grid while preserving its predictive power? The best allocation of a network of a given size on a bounded domain is an issue widely discussed in the literature (see [3][4]). Methods have been developed in the framework of space-filling designs (see [9][10]), according to some optimality property of experimental designs (see also [11]) or, if the spatial covariogram is known, by minimizing the mean squared prediction error of the response variable using the corresponding kriging model.

In general, given a network defined according to some criteria, two are the issues we have to face with. First, how to find the subgrid maximally representative of both the sample space and the assigned grid. Additionally, the procedure must be flexible enough to include, if available, expert knowledge about sub regions. For example, in semiconductor processes, engineers know that the production process close to the borders of the wafer is often affected by less precision than the one placed at the centre of the wafer. Hence, they suggest, whereas possible, to oversample the former part of the wafer.

The goal is then twofold: the grid representativeness, i.e. how much the new grid represents the starting network, and the spatial coverage, i.e. how much the selected subgrid is spatially spread and able to represent the sample space.

3 Methods in the literature

In [1] the use of the simulated annealing for spatial sampling is discussed.

Let $\Im$ be the family of all possible subsets of size n which can be formed by a finite set of M original points, let S be an element of $\Im$ and Φ: $\Im \rightarrow$P+ a positive function, e.g. the kriging prediction error [5], called the fitness function to be optimized over $\Im$.

Simulating annealing (SA) is a heuristic optimization method for solving the above mentioned problem when the cardinality of $\Im$ is high and its elements cannot be evaluated by enumeration [6]. Given the pair S,S'$\in \Im$, S is preferred to S', say $S \succ S'$, if $\Phi(S)<\Phi(S')$. If $\Phi(S) = \Phi(S')$, then S and S' will be equivalent to each other, S~S'. The algorithm starts from a configuration of points, say $S_0 \in \Im$ and sequentially updates it. At each step i, the current configuration S_i is modified by replacing one point of S_i by one point included in its complement. The candidate point for replacement is selected randomly in a certain neighbourhood of the current sample and it is accepted if this determines an improvement in the value of the fitness function. Otherwise, the decision whether the candidate should be included or not is randomized. The solution space is the set of all the possible reduced sampling grids obtained from the original one and a neighbor of the current state is the grid that differ from it only for one point. The procedure is iterated until the value of the fitness gets stable and cannot be further reduced. The criterion used for updating the current sample, known as the Metropolis criterion, is

$$P(S_i \to S_{i+1}) = 1 \qquad \text{if } \Phi(S_{i+1}) \le \Phi(S_i)$$
$$P(S_i \to S_{i+1}) = exp\left(\frac{\Phi(S_i) - \Phi(S_{i+1})}{c}\right) \qquad \text{if } \Phi(S_{i+1}) > \Phi(S_i)$$

where c is a constant which decreases in value as the algorithm keeps going, lowering the probability of accepting less favourable configurations. The method showed quite good capability to reproduce closely the response surface estimated using the full grid.

In [2] a geometric approach is adopted.
Suppose to allocate with no constraints a grid of size n<M according to a pre-chosen scheme, e.g. according to a design with some optimal properties. Let A_n be a measure of accuracy, i.e. how much the selected grid is close to the desired one. The lower A_n is the better. Let R_n be a measure of representativeness, i.e. how much the selected grid is close to the original one. Clearly the closer the selected is to the original one the better the points' allocation is. We can balance representativeness and accuracy through $[\lambda A_n + (1-\lambda)R_n]$ where λ is a weight that may be subjectively assigned according to the relevance of R with respect to A. To consider the capability of the selected new grid to cover adequately the whole surface the entropy, S_n, computed using the areas of the corresponding Voronoi tasselation is considered. The larger the entropy value is the better the selected grid covers the wafer area. By aggregating all the information reported above, the final objective function was to search for the size n such that

$$\max S_n \quad s.t. \quad \min_n : [\lambda A_n + (1-\lambda)R_n]$$

The method allows to explore subgrids that in some manner are sub optimal images of optimal designs and that allow to represent the whole region.

4 Extensions

One of the main drawbacks of the previous approaches is the computational time needed to find the optimal solution. A possible extension is the Tabu Search [7]. Tabu search is an heuristic technique for solving a combinatorial optimization problem iteratively that moves from one admissible solution to another one based on the values of the objective function. The algorithm locally searches for t solutions to the problem by exploring the solution space by overcoming the problem to stop at local optimalities. The technique is guided by memory structures to prevent the algorithm from repeatedly visiting the same solutions, ensuring greater efficiency and

less computation time. At each iteration a neighbour is considered from which some solutions are removed and not reachable from successive iterations (tabu configurations). Because of the efficient computational time need by this approach it can be used for multiobjective problems. Relevant cases are the production processes where multivariate responses must be considered. In these cases a co-kriging approach can be adopted. For example if the response is a bivariate variable the objective function will be of the form

$$P = \lambda * Obj_1 + (1-\lambda)*Obj_2$$

with $0<\lambda<1$, and Obj_i, for i=1,2, the prediction error of the response variables.

References

1. Borgoni, R., Radaelli, L., Tritto, V., Zappa D.: Optimal reduction of a spatial monitoring grid: Proposals and applications in process control. Computational Statistics & Data Analysis **58,** 407-419 (2013)
2. Borgoni, R., Zappa D.: Selecting subgrids from a spatial monitoring network: Proposal and application in semiconductor manufactoring process, Qual. Reliab. Eng. International **33**, 1249—1261 (2017)
3. Lekivetz, R., Jones, B.: Fast Flexible Space-Filling Designs for Nonrectangular Regions, Qual. Reliab. Eng. International **31**, 829–837 (2015)
4. Pronzato, L., Müller, W.G.: Design of computer experiments: space filling and beyond, Statistics and Computing **22**, 681-701, (2012)
5. Van Groenigen, J.W., Siderius, W., Stein, A.: Constrained optimization of soil sampling for minimization of the kriging variance, Geoderma **87**, 239-259 (1999)
6. Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines - A Stochastic Approach to Combinatorial Optimization and Neural Computing. Wiley, New York (1990)
7. Goldberg, D. E.: Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley (1989)

Extension of the JCGM 106:2012 - Conformity assessment of multicomponent items and finite statistical samples

Estensione del JCGM 106:2012 - Valutazione di conformità di oggetti multicomponente e campioni di numerosità finita

Francesca Pennecchi and Ilya Kuselman

Abstract The JCGM 106:2012 document provides guidelines on how to perform conformity assessment of a (scalar) property of interest of a single item (a product, material, object, etc.). In particular, based on a Bayesian approach, it indicates how to model and calculate specific and global risks of the consumer and the producer. In the present work, the JCGM 106 approach is generalized to items that are multicomponent materials (each component having its own property that should undergo conformity assessment with respect to its own requirements), and to a set of *N* items drawn from a common population (the probability of having a certain number of conforming items within this sample needs to be calculated).

Abstract *Il documento JCGM 106:2012 fornisce indicazioni su come effettuare la valutazione di conformità di una grandezza (scalare) di interesse, relativa ad una singola entità (un prodotto, materiale, oggetto, ecc.). In particolare, basandosi su un'impostazione bayesiana, il documento spiega come modellizzare e calcolare i rischi specifici e globali del consumatore e del produttore. In questo lavoro, la metodologia viene generalizzata ad oggetti multicomponente (in cui, per ciascuna componente, la relativa grandezza di interesse deve essere valutata se conforme o meno ai rispettivi requisiti), e per un insieme di N oggetti estratti da una stessa popolazione (occorre calcolare la probabilità che in quel campione ci sia un certo numero desiderato di oggetti conformi).*

Key words: conformity assessment, consumer's and producer's risk, multicomponent items, finite sample.

[1] Francesca Pennecchi, Istituto Nazionale di Ricerca Metrologica (INRIM); email: f.pennecchi@inrim.it

Ilya Kuselman, Independent Consultant on Metrology, 4/6 Yarehim St., 7176419 Modiin, Israel; email: ilya.kuselman@bezeqint.net

Francesca Pennecchi and Ilya Kuselman

1 The JCGM 106:2012 approach for conformity assessment

According to the definition of the JCGM 106:2012 document [5] "conformity assessment is any activity undertaken to determine, directly or indirectly, whether a product, process, system, person or body meets relevant standards and fulfils specified requirements". The document provides guidance and procedures for assessing the conformity of an item (entity, object or system).

The conformity assessment (CA) of an item of interest, such as a gauge block of an industrial production or a sample of air from an environment under air quality control, requires to check whether a certain property of interest of the item, i.e., the measurand [6, Sec. 2.3] (e.g., the length of the gauge block or the concentration of a specific pollutant within the air sample), lies within a prescribed tolerance interval (TI). In general, however, the true value η of the measurand is never completely known but it needs to be measured. Hence, CA decisions such as "the item is conforming" or "the item is rejected" rely on a measured value η_m, which has always a measurement uncertainty (MU) [4] associated with it. The accept/reject decision is based on the evidence of η_m falling or not, respectively, in an acceptance interval (AI) of permissible measured values. AI can differ from TI, in a way to favour either the consumer's or the producer's interests, and is typically established by taking into account the value of MU associated with η_m.

In order to use all available knowledge on the measurand, a Bayesian modelling is considered for the measurable quantity Y: the pre-measurement information is represented by a prior pdf $g_0(\eta)$, whereas the post-measurement state of knowledge is modelled by the posterior pdf $g(\eta|\, \eta_m)$, which is given by the following expression:

$$g(\eta|\, \eta_m) = C\, g_0(\eta)\, h(\eta_m|\, \eta), \tag{1}$$

where C is a normalizing constant and $h(\eta_m|\, \eta)$ is the likelihood function of η given η_m, that is the pdf of possible η_m values of the measuring system output quantity Y_m, at the true value $Y = \eta$ of the measurand.

Based on eq. (1), the following risks of erroneous decisions can be defined and calculated for both the consumer (probability of accepting the item, when it should have been rejected) and the producer (probability of falsely rejecting the item), respectively:

- Specific risks (for a specific item)

$$R_c^* = \int_{TI'} g(\eta|\, \eta_m)\, d\eta \text{ for a specific } \eta_m \text{ in AI}, \tag{2}$$

$$R_p^* = \int_{TI} g(\eta|\, \eta_m)\, d\eta \text{ for a specific } \eta_m \text{ in AI'}. \tag{3}$$

- Global risks (for an item to be chosen at random from the production process)

$$R_c = \int_{TI'} \int_{AI} g_0(\eta)\, h(\eta_m|\, \eta)\, d\eta_m\, d\eta, \tag{4}$$

$$R_p = \int_{TI} \int_{AI'} g_0(\eta)\, h(\eta_m|\, \eta)\, d\eta_m\, d\eta. \tag{5}$$

In eqs. (2-5), TI' and AI' indicate the set of true and measured values which lies outside TI and AI, respectively. Eqs. (2-3) involve integration of the posterior pdf (1),

whereas eqs. (4-5) are double integrals of the joint pdf $f(\eta, \eta_m) = g_0(\eta)\ h(\eta_m|\ \eta)$ of variables Y and Y_m.

2 Generalization to multicomponent items

The JCGM 106 "deals with items having a single scalar property with a requirement given by one or two tolerance limits" and states: "the concepts presented can be extended to more general decision problems". For example, when, for each item, more than one measurable quantity should undergo CA (like in the case of several properties of a blood sample in a routine blood analysis), the CA would be performed separately for every parameter of interest. However, when CA for each particular component is successful and particular consumer and producer's risks (2-5) are acceptable, the total probability of a false decision on the conformity of the material as a whole might still be significant.

The IUPAC projects [1, 2] and corresponding IUPAC/CITAC Guide [7], addressed this topic by defining and modelling total consumer's risks and producer's risks (both specific and global):

- $R^*_{total(c)}$ is the probability that a specific accepted item[1] does not conform, as a whole, i.e., the true value of at least one component is not conforming;
- $R^*_{total(p)}$ is the probability that the true values of all components in a specific rejected item[2] are conforming;
- $R_{total(c)}$ is the probability that an item with a non-conforming true value of one or more components will be accepted based on a statistical analysis of performed measurement results;
- $R_{total(p)}$ is the probability that an item with conforming true values of all the components will be rejected based on a statistical analysis of performed measurement results.

The current project [3] "Influence of a mass balance constraint on uncertainty of test results of a substance or material and risks in its conformity assessment", is tackling the CA of compositional (multicomponent) items, whose components are linked by a mass balance constraint.

2.1 Total risks for independent variables

[1] A multicomponent item is accepted if the measured value of each component lays in its own acceptance interval.

[2] A multicomponent item is rejected when the measured value of at least one of the components lays outside its own acceptance interval.

When the measurable quantities Y_i and the measuring system output quantities Y_{im} are independent, component by component i, it can be demonstrated, based on the law of total probability, that the total specific risks $R^*_{total(c,p)}$ are a combination of particular specific ones (2) or (3), respectively [7]. For example, for just two components under CA, $R^*_{total(c)} = R_{1c}{}^* + R_{2c}{}^* - R_{1c}{}^* R_{2c}{}^*$. Total global risks $R_{total(c,p)}$ result instead in a combination of particular global risks (4) or (5), weighted by probabilities $P(C_i) = P(Y_{im} \text{ in } AI_i)$ [7]. For two components under CA, for example, one has $R_{total(c)} = P(C_2)\, R_{1c} + P(C_1)\, R_{2c} - R_{1c}\, R_{2c}$.

A case study on the monitoring total suspended particulate matter (TSPM) in ambient air, where pollutant concentrations caused by three stone quarries were taken as independent, showed a total global risk higher than the three particular ones.

2.2 Total risks for correlated variables

When correlations are present among measurable quantities Y_i and/or measuring system output quantities Y_{im}, for $i = 1, \ldots, n$, a multivariate Bayesian approach is adopted, involving multivariate pdfs and likelihood functions:

$$g(\boldsymbol{\eta}|\, \boldsymbol{\eta}_m) = C\, g_0(\boldsymbol{\eta})\, h(\boldsymbol{\eta}_m|\, \boldsymbol{\eta}), \tag{6}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_m$ are the vectors of true and measured values of the components, respectively. In this case, the total specific risks are [7]:

$$R^*_{total(c)} = 1 - \int_{\mathbf{TI}} g(\boldsymbol{\eta}|\, \boldsymbol{\eta}_m)\, d\boldsymbol{\eta} \quad \text{for a specific } \boldsymbol{\eta}_m \text{ in } \mathbf{AI}, \tag{7}$$

$$R^*_{total(p)} = \int_{TI1} \ldots \int_{TIv} \int_{R} \ldots \int_{R} g(\boldsymbol{\eta}|\, \boldsymbol{\eta}_m)\, d\boldsymbol{\eta} \quad \text{for a specific } \boldsymbol{\eta}_m \text{ outside } \mathbf{AI}, \tag{8}$$

where $\mathbf{AI} = [AI_1 \times \ldots \times AI_n]$, $\mathbf{TI} = [TI_1 \times \ldots \times TI_n]$, the integral in eq. (7) is a multiple one, and $\boldsymbol{\eta}_m$ in eq. (8) is outside **AI** at the first $v \leq n$ components.

The total global risks are [7]:

$$R_{total(c)} = \int_{\mathbf{TI}'} \int_{\mathbf{AI}} g_0(\boldsymbol{\eta})\, h(\boldsymbol{\eta}_m|\, \boldsymbol{\eta})\, d\boldsymbol{\eta}_m\, d\boldsymbol{\eta}, \tag{9}$$

$$R_{total(p)} = \int_{\mathbf{TI}} \int_{\mathbf{AI}'} g_0(\boldsymbol{\eta})\, h(\boldsymbol{\eta}_m|\, \boldsymbol{\eta})\, d\boldsymbol{\eta}_m\, d\boldsymbol{\eta}, \tag{10}$$

where the multiple integration with respect to $\boldsymbol{\eta}$ on **TI'** in eq. (9) addresses all those cases in which at least one true value η_i is outside its TI_i, whereas the multiple integration with respect to $\boldsymbol{\eta}_m$ on **AI'** in eq. (10) addresses all those cases in which at least one measured value η_{im} is outside its AI_i.

A case study on CA of a four-component alloy showed the impact of correlation among the components on the total risk: neglecting correlations would lead to an overestimation of the global consumer risk.

2.3 Total risks for compositional data

Extension of the JCGM 106:2012 - Conformity assessment of multicomponent items and finite statistical samples

When the components of a multicomponent item are subject to a mass balance constraint, e.g. $\Sigma\eta_i$ = 100 %, they are intrinsically correlated. A so-called 'spurious' correlation is then observed in addition to other possible natural and/or technological correlations. Moreover, when choosing an appropriate prior pdf, the constraint for the true values of each component to lay in the domain [0, 100] % has to be taken into account. An approach based on Monte Carlo simulations from a multivariate truncated normal pdf followed by a closure operation was applied to a case study on the CA of a specific sausage product, made of four components (fat, protein, moisture, salt) [8].

3 Extension to a finite sample of items

A further direction toward which the JCGM 106 framework could be extended is the CA of a finite sample of *N* items drawn from a common population ("The concepts presented can be extended to more general conformity assessment problems based on measurements of a set of scalar measurands" [5]). The idea is related to CA of a sample of *N* units from a population, e.g., a batch of *N* items from a population of batches at a factory, producing such batches continuously, where each item should be tested (the item parameters are to be measured). This may be necessary in an aircraft, military or car industry, in clinical analysis of a group from a population (schoolers of a specific school, bus drivers of a specific company, chemists of a laboratory, etc), in assessing the results of air monitoring in a specific region, etc.

Therefore, a recent research activity aims at generalization of specific and global risks (2-5) for a sample of items, that is, at answering the following two questions, respectively:

1) Given a sample of *N* measured items (each characterized by specific risks (2-3)), among which *K* have been measured within their AI (good measured values - GMV), which is the probability that at least *J* of the *N* corresponding true values were actually conforming – or, equivalently, the probability that less than *J* true values were non-conforming (bad true values - BTV)?
2) Considering to randomly drawing a sample of *N* items from a population already characterized by global risks (4-5), which is the probability to have, among them, exactly K_1 that are GMV&BTV, K_2 that are BMV&GTV, K_3 that are GMV&GTV and K_4 that are BMV&BTV? Notice that K1 + K_2 + K_3 + K_4 = *N*.

3.1 Specific risk

In order to answer question 1) above, we can resort to a discrete random variable (r.v.) V counting, among N, how many of the measured values η_{im} (where $i = 1, \ldots, N$ is now the index enumerating the items in the sample) actually come from a corresponding good true value η_i. V is then the sum of N independent Bernoulli r.v., each with its own success probability P(Y_i in TI | η_{im}), which is equal to 1 - R_{ic}*, if η_{im} is a GMV, or to R_{ip}*, if η_{im} is a BMV.

Therefore, considering question 1):

- $V \sim$ Poisson binomial(1 - R_{1c}*, …,1 - R_{Kc}*, $R_{(K+1)p}$*,…, R_{Np}*),
- and the answer is given by P($V \geq J$) (which is also equal to 1 - P($V < J$)).

Setting the desired[1] probability P($V \geq J$) leads to the solution of the inverse problem "which is the maximum J value allowing to reach the desired probability to actually have at least J good true values in that sample?".

3.2 Global risk

In order to answer question 2) in Sec. 3, let us consider that P(GMV&BTV) = R_c and P(BMV&GTV) = R_p, by definition, whereas

$$\mathrm{P(GMV\&GTV)} = \int_{TI} \int_{AI} g_0(\eta)\, h(\eta_m|\, \eta)\, d\eta_m\, d\eta, \quad (11)$$

$$\mathrm{P(BMV\&BTV)} = \int_{TI'} \int_{AI'} g_0(\eta)\, h(\eta_m|\, \eta)\, d\eta_m\, d\eta. \quad (12)$$

Expressions (11) and (12) provide, in terms of a confusion matrix notation, the (probability of) true positives (p_{TP}) and true negatives (p_{TN}), respectively. Therefore, the discrete r.v. W able to answer question 2) has a multinomial pmf with parameters N and the probabilities provided by eqs. (4-5) and (11-12):

- $W \sim$ multinomial(N; R_c, R_p, p_{TP}, p_{TN}) (Note that $R_c + R_p + p_{TP} + p_{TN} = 1$),
- and the answer is given by P($W = [K_1, K_2, K_3, K_4]$).

References

1. IUPAC Project 2016-007-1-500 (https://iupac.org/project/2016-007-1-500)
2. IUPAC Project 2018-004-1-500 (https://iupac.org/project/2018-004-1-500)
3. IUPAC Project 2019-012-1-500 (https://iupac.org/project/2019-012-1-500)
4. JCGM-WG1: JCGM 100:2008 - Evaluation of measurement data – Guide to the expression of uncertainty in measurement (https://www.bipm.org/en/committees/jc/jcgm/publications)
5. JCGM-WG1: JCGM 106:2012 - Evaluation of measurement data – The role of measurement uncertainty in conformity assessment (https://www.bipm.org/en/committees/jc/jcgm/publications)
6. JCGM-WG2: JCGM 200:2012 - International vocabulary of metrology – Basic and general concepts and associated terms (VIM) (https://www.bipm.org/en/committees/jc/jcgm/publications)
7. Kuselman, I., Pennecchi, F. R., da Silva, R. B., Hibbert, D. B.: IUPAC/CITAC Guide: Evaluation of risks of false decisions in conformity assessment of a multicomponent material or object due to measurement uncertainty (IUPAC Technical Report). Pure Appl. Chem. **93**(1), 113–154 (2021)

[1] A requirement might be, for example, to reach at least 90 % probability of having at least 90 % of conforming items in the sample.

Extension of the JCGM 106:2012 - Conformity assessment of multicomponent items and finite statistical samples

8. Pennecchi, F. R., Kuselman I., Di Rocco, A., Brynn Hibbert, D. B., Semenova, A. A.: Risks in a sausage conformity assessment due to measurement uncertainty, correlation and mass balance constraint. Food Control **125**, 107949 (2021)

Session of solicited contributes SS5 – *Big Data and Large-dimensional Data*
Organizer and Chair: Stefania Mignani

ROBOUT: a conditional outlier detection methodology for large-dimensional data

ROBOUT: una metodologia per identificare i valori anomali condizionati su dati di grandi dimensioni

Matteo Farnè and Angelos Vouldis

Abstract This paper presents a fast methodology, called ROBOUT, to identify outliers conditional on a set of linearly related predictors, retrieved from a large granular dataset. ROBOUT is shown to be effective and particularly versatile compared to existing methods in the presence of various data idiosyncratic features. Specifically, ROBOUT can identify observations with outlying conditional variance when the dataset contains element-wise sparse variables, multicollinearity and large number of variables compared to the number of observations. ROBOUT entails a robust selection stage of the statistically relevant predictors (by using a Huber or a quantile loss), the estimation of a robust regression model based on the selected predictors (by LTS or MM), and a criterion to identify conditional outliers based on a robust measure of the residuals' dispersion. The methodology is also applied to a granular supervisory banking dataset collected by the European Central Bank.

Abstract *Questo articolo presenta una metodologia rapida, chiamata ROBOUT, per identificare valori anomali condizionatamente a un insieme di previsori linearmente correlati, ritrovati da un grande dataset granulare. ROBOUT si mostra essere efficace e particolarmente versatile rispetto ai metodi esistenti in presenza di un numero di particolari caratteristiche dei dati. In particolare, ROBOUT riesce a identificare osservazioni con una varianza condizionata anomala quando il dataset contiene variabili sparse, multicollinearità, e un grande numero di variabili rispetto alle osservazioni. ROBOUT contiene un passo di selezione robusta dei previsori statisticamente rilevanti (tramite perdita di Huber o del quantile), la stima robusta di un modello di regressione (tramite LTS o MM) e un criterio per identificare i valori anomali condizionati rispetto a una misura robusta della dispersione dei residui. La metodologia è applicata a un dataset granulare di supervisione della Banca Centrale Europea.*

Key words: conditional outlier, robust regression, variable selection, large dimension, sparsity

[1] Matteo Farnè, Department of Statistical Sciences, University of Bologna; email: matteo.farne@unibo.it

Angelos Vouldis, European Central Bank, Directorate General Statistics; email: angelos.vouldis@ecb.europa.eu

Matteo Farnè and Angelos Vouldis

1 Introduction

Data quality is a fundamental prerequisite for any kind of quantitative analysis, and the large datasets which are becoming increasingly available present specific challenges to the task of monitoring and ensuring data quality. One critical aspect of the data quality monitoring is outlier detection, i.e., the identification of values which are either obviously mistaken or seem to be unjustified from an empirical perspective.

In this paper, we focus on outlier detection in large-dimensional datasets, where the number of variables p (i.e., the dimension of the data space) and the number of observations n (i.e., the sample size) are large. The dataset may also be 'fat', i.e., featuring $p > n$. Such datasets arise in diverse fields such as bioinformatics, economics, neuroscience, signal processing and others. For instance, the supervisory banking dataset analyzed in this paper contains 453 variables for 365 banks. Our aim is to retrieve from such a dataset any anomaly in a target variable y with respect to a set of K related variables (the predictors of y) that constitute a subset of the $p \gg K$ variables of the dataset (the candidate predictors of y). The best predictors of y are ex-ante unknown and need to be identified by the outlier detection algorithm. Furthermore, we assume that outlier identification for variable y takes place in a challenging environment in which the remaining $p - K$ variables may be elementwise sparse, a feature that increases correlation across the variables of the dataset and may derail some of the existing outlier detection algorithms.

Let us formalize the described problem in probabilistic terms. In [1], unconditional outliers are defined as instances which fall into a low probability density region of $f(y)$, where $f(y)$ is the unconditional probability density function of the scalar target variable y. Instead, conditional outliers are defined as instances of y which fall into a low probability density region of $f(y|\mathbf{x}) = f(y, \mathbf{x})/f(\mathbf{x})$, where $f(y|\mathbf{x})$ is the conditional probability density function of the scalar target variable y given a vector of related variables $\mathbf{x}$. In this paper, we assume that the expected value of $f(y|\mathbf{x})$ is a linear function of the variables in $\mathbf{x}$. Note that this does not constrain the nature of the prescribed relationships between y and $\mathbf{x}$, because we can always include quadratic and exponential functions of specific variables in the vector $\mathbf{x}$.

For the sake of simplicity, let us suppose that for the single observation $i \in \{1, \cdots, n\}$ it holds $y_i|\mathbf{x_{D,i}} \sim N(a_i + \mathbf{x'_{D,i}}\boldsymbol{\beta}, {\sigma_i}^2)$, where y_i is the value of y for unit i, $\mathbf{x_{D,i}}$ is a $K \times 1$ vector of predictors for unit i, $\boldsymbol{\beta}$ is a $K \times 1$ vector of regression coefficients and ${\sigma_i}^2$ is the variance of y_i conditional on $\mathbf{x_{D,i}}$. We define a set of outlier indices O with $O \subseteq \{1, \cdots, n\}$ such that $|O| = \alpha n$. The parameter α represents the contamination rate of the dataset and $\alpha \in [0, 0.5]$. We assume, without loss of generality, that for any $i \notin O$ it holds $a_i = a$ and ${\sigma_i}^2 = \sigma^2$. We denote by D the set of predictor indices, such that $|D| = K$ and $D \subseteq \{1, \cdots, p\}$, and by D' the complementary set of D with respect to the set $\{1, \cdots, p\}$. We also assume that $\mathbf{x_{D,i}} \sim N(\boldsymbol{\mu}_i, \mathbf{I}_K)$ with $\boldsymbol{\mu}_i = \mathbf{0}_K$ for any $i \notin O$, without loss of generality. The $n \times K$ matrix $\mathbf{X_D}$ contains as

ROBOUT
columns the predictor variables, indexed by D, and the $n \times (p - K)$ matrix $\mathbf{X_{D'}}$ contains as columns the non-predictor variables, indexed by D'.

Definition 1 A *conditional outlier in mean* is defined as any observation $i \in O$ such that y_i falls in a low probability density region of $N(a + \mathbf{X_{D,i}}\boldsymbol{\beta}, \sigma^2)$ because $a_i \gg a$ or $a_i \ll a$.

Definition 2 A *conditional outlier in variance* is defined as any observation $i \in O$ such that y_i falls in a low probability density region of $N(a + \mathbf{X_{D,i}}\boldsymbol{\beta}, \sigma^2)$ because ${\sigma_i}^2 \gg \sigma^2$.

Definition 3 The matrix $\mathbf{X_D}$, is defined as γ-sparse if it contains $\gamma \times n \times (p - K)$ randomly positioned zeros, with $\gamma \in [0,1)$.

Practically speaking, Definitions 1 and 2 may represent in y measurement errors, idiosyncratic events, unpredictable shocks etc., as well as structural differences in the data generating mechanism. Definition 3 represents the case where the non-predictors in $\mathbf{X_D}$, contain missing or empty values.

In this paper, we present a versatile method for conditional outlier detection, called ROBOUT, which is efficient as regards computational cost (i.e., proportional to $O(pn)$) while it returns a reliable solution to conditional outlier recovery for datasets featuring diverse statistical properties. The proposal of ROBOUT improves existing methods in the literature by providing improvement in both outlier detection performance and computational speed while in addition being robust to various dataset particularities such as the presence of multicollinearity, γ-sparsity or large number of variables compared to the number of observations. We conduct an extensive simulation study that includes scenarios featuring various levels of data sparsity and outlier proportions, and different relative data dimensions (cases of both $p>n$ and $p<n$). The results show that ROBOUT performs better than the existing alternatives especially when we consider the overall ex ante performance, i.e., when the statistical properties of the dataset are not known, therefore it is a versatile method that is shown to be efficient and effective in this broad range of situations, thus providing a feasible and reliable answer to conditional outlier detection in large dimensions.

2 Data generating process

Let us assume n numerical observations of one response variable y and p additional variables. We call the unknown set of outlier indices O with $|O| = [\alpha n]$ and $\alpha \in [0,0.5]$. The response variable vector $\mathbf{y}$ may be expressed in terms of the following regression model

$$\mathbf{y} = \mathbf{a} + \mathbf{X_D}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

(1)

where $\mathbf{y}$ is the $n \times 1$ vector of the response variable, $\mathbf{a}$ is the $n \times 1$ vector of intercepts, $\mathbf{X_D}$ is the $n \times K$ matrix of predictors (whose columns are indexed by D), $\boldsymbol{\beta}$ is the $K \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of residuals. The same regression model for the single observation $i \in \{1, \cdots, n\}$ can be written as

$$y_i = a_i + \mathbf{x'_{D,i}}\boldsymbol{\beta} + \varepsilon_i$$

(2)

where a_i denotes the intercept for unit i and $\mathbf{x_{D,i}}$ denotes the $K \times 1$ vector of predictors for unit i.

For each non-outlier index $i \notin O$, we assume with no loss of generality that $a_i = a$, $\varepsilon_i \sim N(0, \sigma^2)$, $\mathbf{x_{D,i}} \sim N(\mathbf{0}_K, \mathbf{I}_K)$ and $\mathbf{x_{D',i}} \sim N(\mathbf{0}_{p-K}, \mathbf{I_{p-K}})$, where D' stores the indices of non-predictors. For each outlier index $i \in O$, we assume that outliers can be generated as follows:

- consistently with Definition 1, conditional outliers in mean are generated by assuming $a_i = ma$, with $m > 1$;
- consistently with Definition 2, conditional outliers in variance are generated by assuming $\varepsilon_i \sim N(0, m\sigma^2)$, with $m > 1$.

In addition, consistently with Definition 3, the $n \times (p - K)$ matrix of non-predictors $\mathbf{X_{D'}}$ may be assumed as γ-sparse with $\gamma \in [0,1)$. More, we may allow for multicollinearity by setting $COV(\mathbf{X}_{j^*}, \mathbf{X}_{j^{**}}) = \rho$, $\rho \in [0,1)$, $\forall\, j^*, j^{**} \in \{1, \cdots, p\}, j^* \neq j^{**}$, where $\mathbf{X} = [\mathbf{X_D} \mid \mathbf{X_{D'}}]$ is the complete $n \times p$ data matrix.

3 ROBOUT methodology

ROBOUT methodology for conditional outlier detection is comprised of three steps:

1) Selection of predictors, by solving the following optimization problem:

$$\min_{\{\beta_j\}_{j=0,1,\ldots,p}} \sum_{i=1}^{n} \left(\rho_H(\varepsilon_i) \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

(3)

ROBOUT

where $\varepsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}$, $\rho_H(t) = I(t > \tau)\left(|t| - \frac{\tau}{2}\right) + I(t \leq \tau)\frac{t^2}{2\tau}$, $t \in R$, is the Huber weight function, and λ is a penalization parameter (henceforth SNCD-H objective function). Eq. (3) estimates an elastic-net penalized Huber loss regression, optimized by using the semi-smooth Newton coordinate descent algorithm presented in [2]. The tuning parameter τ is by default posed equal to $IQR(y)/10$. Weighting observations is precisely what renders the results robust in the face of perturbed conditions because it annihilates the influence of conditional outliers. The other robust alternative that we consider substitutes $\rho_H(\varepsilon_i)$ in (3) with the median loss $\rho_Q(\varepsilon_i)$, where $\rho_Q(t) = t\left\{\frac{1}{2} - I(t < 0)\right\}$, $t \in R$ (henceforth SNCD-Q objective function). SNCD-Q estimates an elastic-net penalized median loss regression, optimized in the same way.

2) Robust regression using the predictors identified by the previous step, by employing

- least trimmed squares (LTS) estimation [3], which identifies the $100 \times (1 - \alpha)\%$ most concentrated observations and estimates the regression coefficients on those via ordinary least squares.
- MM-estimation [4], which is a 3-stage procedure, based on M-estimation, minimizing a Huber function of the residuals, and using a robust initialization of the coefficients $\boldsymbol{\beta}$ and the residual scale σ.

3) Outlier detection, by comparing robustly rescaled residuals to the standard normal distribution.

4 A banking data example

In this section, we apply the proposed conditional outlier detection procedure, ROBOUT, to a real dataset that contains granular data on the activities of the largest euro area banks, both on the asset and the liability sides of their balance sheet. These data are submitted by the European banks to the European Central Bank in the context of their supervisory reporting requirements. Our sample consists of a cross section of n=365 banks and p=453 variables. The reference date of the data is end-2014.

Looking at the regression models estimated via the SNCD-H+MM and SNCD-Q+MM options, we can judge that $K = 5$ is the most appropriate choice, since the estimated coefficients are all strongly significant (see Table 1), and their signs are meaningful. We can see that derivatives (both for trading and hedging purposes) at notional amount and debt securities at amortised cost have a positive impact on the log-assets, while hedging derivatives at carrying amount and loans and advances on demand and short notice to credit institutions have a negative impact. The adjusted R-squared overcomes 82%.

Table 1: *Banking data: estimated SNCD-H+MM/ SNCD-Q+MM regression output, with number of regressors K=5.*

Identifier	Name	Estimate	Standard error	t-value	p-value
	Intercept	18.9432	0.1564	121.115	<2E-16
{'F0101_r240_c010'}	Derivatives – Hedge accounting – Carrying amount	-43.0174	13.9705	-3.079	0.00224
{'F0500_r010_c030'}	Loans and advances on demand [call] and short notice [current account] – Credit institutions	-4.7105	0.9455	-4.982	9.81E-07
{'F0801a_r360_c030'}	Debt securities issued – Amortised cost	4.3541	1.2376	3.518	0.00049
{'F1000_r320_c030'}	Derivatives: Trading – OTC rest – Notional amount – Total trading	7.8773	0.853	9.235	<2E-16
{'F1101_r500_c030'}	Derivatives – Hedge accounting – Notional amount – Total Hedging	4.2476	0.7814	5.436	1.01E-07
Robust residual standard error:	1.52				
Multiple R-squared:	0.8241				
Adjusted R-squared:	0.8216				

The nature of recovered conditional outliers can be understood by looking at their relative standing in the distribution of recovered predictors. In Table 7, we can observe that the recovered conditional outliers present outstanding values of derivatives (both for trading and hedging) at notional amount. Therefore, the set of recovered outliers mainly identifies small banks whose small size is an outlier when compared to the relatively large amount of trading derivatives present in their balance sheet, thus uncovering an unexpected outlyingness dimension for euro area banks.

ROBOUT

Table 2: *Median and median absolute deviation (MAD) of log-assets and the five selected predictors for outliers and non-outliers.*

		log – size	Derivatives – Hedge accounting – Carrying amount	Loans and advances on demand – Credit institutions	Debt securities issued – Amortised cost	Derivatives: Trading – OTC rest – Total trading	Derivatives – Hedge accounting Total Hedging
Non – outliers	median	19.8104	0.0002	0.0132	0.0387	0.0078	0.0481
Outliers	median	19.9485	0.0023	0.0135	0.1046	0.5608	0.1498
Non – outliers	MAD	1.7312	0.0067	0.0562	0.0872	0.1036	0.1601
Outliers	MAD	1.6705	0.0044	0.0369	0.121	0.3941	0.3388

5 Conclusions

In this paper, we propose a new conditional outlier detection methodology, called ROBOUT. ROBOUT is very versatile and flexible, as it can robustly spot conditional outliers under many different perturbed conditions and different combinations of sample size and dimension. In addition, ROBOUT works efficiently on datasets with many observations and variables as it is computationally lighter than alternative methods, and it can robustly select the most relevant predictors for each analysed variable.

References

1. Hong, C., Bauknecht, M.: Multivariate Conditional Anomaly Detection and Its Clinical Application. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4239--4240. Association for the Advancement of Artificial Intelligence (2015)
2. Yi, C., Huang, J.: Semismooth newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. Journal of Computational and Graphical Statistics **26**(3), 547--557 (2017)
3. Rousseeuw, P.J., Van Driessen, K.: Computing LTS regression for large data sets. Data mining and knowledge discovery **12**(1), 29--45 (2006)
4. Yohai, V.J.: High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, 642--656 (1987)

Behaviors, emotions and opinions in modern citizen or customer relationship systems: a correct integration of small and big data for hyper-targeting, personal advertising and look-alike

Comportamenti, emozioni e opinioni nei moderni sistemi di relazione col cittadino o col cliente: una corretta integrazione degli small data e dei big data per il micro-targeting, la comunicazione personalizzata e il processo look-alike

Furio Camillo

Abstract The paper presents two different case studies of virtuous integration between small data and big data in the field of communication and marketing in hyper-targeting and CRM domains.
For some years now, the debate of marketing experts and more generally of business strategies has been enriched with contributions on the relationship between small data and big data. As is well-known, big data increasingly represents the behavior of individuals by collecting and arranging traces that people leave using various devices that pervade our lives every day: the car, household appliances, the web browsing, use of apps, television decoders, presence on social networks, etc. Precisely because most of the data is generated by default by our devices, sometimes the collection and storage generate a lot of noise in the data: missing data, poorly detected or aberrant data, database with some unclear semantic construct. Contribution of "small" data, whose generation is instead designed, on controlled and not self-selected samples, to detect elements that are really useful for the actions of policy makers, public and private, becomes crucial every time you have a database of "unencrypted individuals" (citizens or customers) and you want to make targeted communication. For example, by designing a psychographic questionnaire on a sample of individuals, some construct of primary interest can be estimated, which will then have to be extended to the rest of the database (look-alike audience extension), useful to project a micro-targeting process for personalized advertising (Quaing, 2016; Severadova, 2019)

Abstract *Il contributo presenta due diversi case studies di integrazione virtuosa fra gli small data e i big data in ambito di comunicazione e marketing.*
Da qualche anno il dibattito degli esperti di marketing e più in generale di strategie di business si è arricchito di contributi autorevoli circa la relazione fra small data e big data. Come è noto, i big data rappresentano sempre di più il comportamento degli individui mediante la raccolta e la sistemazione delle tracce che le persone lasciano usando i vari devices che ogni giorno pervadono la nostra vita: l'autovettura, gli elettrodomestici, il browser, la navigazione sul web, l'uso delle app, il decoder

Furio Camillo, Department of Statistical Sciences, University of Bologna, Bologna, Italy

televisivo, la presenza sui social networks, ecc. Proprio perché la maggior parte dei dati è generata di default dai nostri devices, a volte la raccolta e l'archiviazione generano molto rumore nei dati: dati mancanti, dati mal rilevati o aberranti, dati con un costrutto semantico poco chiaro. Il contributo di dati "small", la cui generazione invece è progettata, su campioni controllati e non autoselezionati, per rilevare elementi davvero utili alle azioni dei policy makers, pubblici e privati, diventa cruciale ogni volte si abbia un database di individui "in chiaro" (cittadini o clienti) e si voglia fare della comunicazioe mirata. Ad esempio progettando un questionario psicografico su un campione di individui potrà essere stimato qualche costrutto di primario interesse, che dovrà poi essere esteso al resto del database e usato per costruire un processo di micro-targeting della comunicazione pubblicitaria (Quaing, 2016).

Key words: personalized advertising, hyper-targeting, psycographic segmentation, extension models, semiometric segmentation, personal values

1 Introduction and history

The advertising systems and the algorithms they use are constantly evolving and expanding the possibilities for reaching potential customers. Hyper-targeting (also called micro-targeting) is the use of detailed customer data and marketing automation to deliver highly targeted and personalized messages across a large number of channels. These campaigns are designed to appeal to specific people or small groups of customers. By using the ability to process large amounts of data through innovations, such as predictive analytics, marketers can gain a deeper understanding of their audiences, focusing on very reactive specific sub-segments and not on the entire segment of interest. Today's highly fragmented, competitive and fast market has contributed to the emergence of new challenges that are reflected in the diversification of consumer habits and needs, in greater corporate exposure, in the growth of customers' information power and in growing competitive pressure. Introduction of these factors has shifted the focus of interest from the product to the customer, defining a customer-centric market, as opposed to a product-centric vision. Transition from the traditional approach to the current one was not immediate: in the early stages of industrialization, the production of goods was planned from above in a standardized way, in order to offer consumers a pre-packaged product with homogeneous characteristics. The homologation of the offer found its reasons in the belief of a substantial "social immobility" (Grassi, 2002), or rather of a static uniformity of the mass population at a behavioral and value level. And it is in this presumed one-dimensional universe of thought that the critique of the philosopher Herbert Marcuse of "one-dimensional" man can be found (Marcuse, 1964). Mass society would therefore be made up of individuals at the mercy of manipulation from above, dehumanized into mere puppets of a puppet theatre. The scientific interest in the cultural dimension, inaugurated in the 1950s by Cultural studies, reveals the importance of the context within the communication process, which finds its concreated application in Jakobson's (1966) communication model. Starting from the limits attributed to the unidirectionality of the cybernetic model of Shannon and

Whaver (1949), which proved reductive to account for communication between human beings, the Russian linguist Roman Jakobson introduces two key factors: the code and the context. In order to be understood by the recipient (addressee), the message must, in fact, be formulated by the sender (addresser) using a code (code) known by both actors, or by a set of signs and rules for combining them. In the scenario described above of the "one-dimensional" man, the decoding phase takes on a crucial role: it allows the mass to be elevated to an active role of selective reading and interpretation of the media products addressed to it. According to the encoding-decoding model of Hall (1973) it is no longer possible to speak of a single compact and homogeneous public, but of a "multitude of audiences" that interpret the same message in a different way according to their culture. Not only the external context, which is reflected in the "encyclopedic competence" (Eco, 1984) of the individual and in interpersonal relationships, has a significant impact on the decoding process, but also the individual conscience, which reveals its nature as a "homo duplex" (Durkheim, 1898), stretched between the two opposite poles of individualism and collectivism. This aspect is emphasized by the exponents of Gestalt psychology, including Köler (1940) and Wertheimer (1945) according to whom the perception of external stimuli is considerably influenced by internal events, such as values, expectations and needs. At the same time, the reception of information is also subjected to an "elaborate control through the past" (Mead, 1934), drawing on the memory and previous experiences of each individual. Therefore, the message is not only subject to selective acquisition, but its interpretation changes according to the type of recipient: "the object reveals itself to the extent that the subject expresses itself" asserts the philosopher Luigi Pareyson (1985) . At this point, it becomes clear that the individual can no longer be considered as a passive receptor of external stimuli, but as an active selector who interprets the message based on the culture in which he is inserted and its internal characteristics. Therefore, the overcoming of the structural asymmetry intrinsic to the conception of mass communication and the rejection of an indistinct message for all types of recipients is decreed.

2 Case study 1: From a qualitative future study to a psychographic questionnaire for a profiling mass-survey

The construction of the questionnaire items was supported by several key concepts of the linguistics and semiotics, with particular reference to semantics. By exploiting the paradigmatic lexical relationships, different items were constructed starting from words between which there are hierarchical relationships (hyperonymy, hyponymy, troponymy) and opposition (antonym). With particular attention to the relationships of antonymia, the aim was to define couples oppositional of contextual or intrinsic binarity between opposites. For their characteristics, opposites, or gradable antonyms (pairs of distinct terms between polar opposites, mapped on a scale of values (fast / slow, high / low)), opposites, and complementaries, or contradictors (complementary antonyms allow you to divide a domain into two non-overlapping and exhaustive halves (true / false, in / out, static / dynamic)) which allow to make due incompatible items, at the two poles of a scale, without however perfectly dividing the domain in two halves, but leaving an unskilled median area, the definition of which is up to the choice individual of the individual respondent. These relationships are through the

graphics device of the semiotic square (Greimas, 1968) which follows the model of structural phonology developed by Trubeckoj. The fundamental opposition is that of opposition which opposes two contradictory seeds on a semantic axis based on the narrative context or intrinsic level. The doubling of the opposition occurs with the negation of the two subcontrary suits.

3 Case study 2: Topics, manipulation and personal values of customers-citizens: the typology of "disheartened-restless"

Using the semiometric approach (Lebart, 2003), it is possible to construct a semi-automated mechanism for hyper-profiled communication. For example, in order to treat the semiometric cluster "disheartened-restless" can use a communication plastic transformation protocol such as the one described in the figure below. Use "seduction" and "provocation" as types of manipulation to bring the recipient-consumer closer to the object of value, following Greimas' theories of the Canonical Narrative Scheme. The technique of seduction would allow satisfying the hedonistic research and the sense of individualism; the provocation would stimulate the sense of domination and criticism intrinsic to this profile. The thematic center of the linguistic system of this profile is the opposition between the semi-CONFIDENCE and CERTAINTY and between BREAKDOWN and TRADITION. The preferred formants are the use of RED, BLUE and BLACK and the following graphic forms.

References

1. Durkheim, E. 1898 Représentations individuelles et représentations collectives, Revue de Métaphysique et de Morale.
2. Eco, U. 1984 Semiotica e filosofia del linguaggio, Torino, Enaudi
3. Grassi, C. 2002 Sociologia della comunicazione, Mondadori Bruno
4. Greimas, A. J. 1966 Sémantique structurale: recherche de methode, Paris: Larousse
5. Hall, S. 1973 Encoding and Decoding in the Television Discourse Birmingham: Centre for Contemporary Cultural Studies
6. Jakobson, R. 1963 Essais de linguistique générale, Paris: ed. de Minuit
7. Köler, W. 1940 Dynamics in psychology, New York, Miveright
8. Lebart, L. et al. 2014 The Semiometric Challenge: Words, Lifestyles and Values
9. Marcuse, H. 1964 One-dimensional Man: Studies in Ideology of Advanced Industrial Society, New York, Routledge
10. Mead, G.H. 1934 Mind, Self, and Society from the Standpoint of a Social Behaviorist, University of Chicago Press: Chicago.
11. Pareyson, L. 1985 Esistenza e persona, Il Melangolo, Genova.
12. Quiang, M. 2016 A Sub-linear, Massive-scale Look-alike Audience Extension System, JMLR: Workshop and Conference Proceedings 16, 2016
13. Severadova T, 2019 Computer Estimation of Customer Similarity With Facebook Lookalikes: Advantages and Disadvantages of Hyper-Targeting, Article in IEEE Access · October 2019 DOI: 10.1109/ACCESS.2019.2948401
14. Shannon, C.; Weaver, W. 1949 The mathematical theory of communication, Univ. of Illinois Press
15. Wertheimer, M. 1959 Productive thinking, New York, Harper & Row

A model for assessing sea environmental quality

Un modello per valutare la qualità ambientale del mare

Ida Camminatiello and Antonio Lucadamo

Abstract In this paper, we aim to study the key factors which affect the environmental quality of the sea. The Fp-ratio, utilized as an indicator of the trophic state of the coastal waters, is classified into three categories. Given the ordinal nature of the response variable, we believe ordinal logistic regression models are the most proper statistic methodology. As the regressors are strongly correlated, the parameters of ordinal logistic regression models cannot be estimated. For solving this problem and, in general, the consequences of multicollinearity, we develop an approach based on principal components. The new approach will be applied for estimating the Fp-ratio.

Abstract *In questo lavoro, ci proponiamo di studiare i fattori chiave che influenzano la qualità ambientale del mare. Il rapporto Fp, utilizzato come indicatore dello stato trofico delle acque costiere, è classificato in tre categorie. Data la natura ordinale della variabile di risposta, riteniamo che i modelli di regressione logistica ordinale (ORL) siano la metodologia statistica più appropriata. Poiché i regressori sono fortemente correlati, i parametri dei modelli di regressione logistica ordinale non possono essere stimati. Per risolvere questo problema e, in generale, le conseguenze della multicollinearità, sviluppiamo un approccio basato sulle componenti principali. Il nuovo approccio sarà applicato per la stima del rapporto Fp.*

Key words: principal component analysis, ordinal logistic regression, multicollinarity, sea quality.

Ida Camminatiello
University of Campania, Capua (CE), Italy, e-mail: ida.camminatiello@unicampania.it

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

1 Introduction

Attention to the environment has been growing more and more in recent years. In this framework, we aim to study the key factors which affect the environmental quality of the sea.
Several studies (Casotti et al., 2000; Mangoni et al., 2013, 2016) utilize the Fp-ratio, proposed by Claustre (1994) as an indicator of the trophic status of the sea. It is defined as the sum of fucoxanthin (diatoms) and peridinin (dinoflag- ellates), divided by the sum of all diagnostic pigments.
In this research, the Fp-ratio, investigated as the response variable, is classified into three categories: from 1 (low) to 3 (high). We aim to evaluate how the Fp-ratio depends on continuous variables: phytoplankton-related variables (size structure and biomass) and physical parameters (salinity and temperature). Given the ordinal nature of the response variable, we believe ordinal logistic regression (OLR) models are the most proper statistic methodology. As the regressors are strongly correlated, the parameters of ordinal logistic regression models cannot be estimated.
For solving the consequences of multicollinearity, the stepwise selection of the predictors can be performed; Bastien, Esposito Vinzi and Tenenhaus (2005) proposed Partial least squares (PLS) ordinal logistic regression. Some authors (Aguilera and Escabias, 2008; Muhammad Nur Aidi and Tuti Purwaningsih, 2013) touched upon the application of the principal components (PCs) without deepening or formalizing the approach (Camminatiello and Lucadamo, 2019).
In this paper, we develop an approach based on PCs, for dealing with multicollinearity in OLR models (second section). In the third section, the approach will be applied for estimating the Fp-ratio. The last section concludes with some remarks and perspectives.

2 Principal component OLR

The most commonly used OLR models are the adjacent-category, the continuation-ratio, and the proportional odds models (Agresti, 1990; Agresti, 2015). In presence of multicollinearity, the estimation of all the OLR models becomes inaccurate because of the need to invert nearsingular and ill-conditioned information matrices. Therefore, the covariates of the OLR models could be substituted by a reduced number of PCs of the regressors.
As it has been stated that the proportional odds is the most frequently used OLR model in practice (Hosmer and Lemeshow, 2000), we develop our approach for it.
Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_q]$ be the $q = min(n, p)$ PCs of the p regressors and Y an ordinal response variable with K categories observed on n statistical units.
The approach can be summarized through the following steps. In the first step, we create the PCs of the regressors which are linear combinations of the original variables. In the second step, we carry out the proportional odds model on the set of

q PCs. In the third step, we select the number of PCs, $a \leq q$, to be retained in the model, according to different tools (Camminatiello, Lucadamo, 2010). Finally, we carry out the proportional odds model on the chosen subset of PCs. The proportional odds model can be expressed in terms of a PCs as:

$$P(Y \leq k) = \frac{exp(\alpha_k + \gamma_1 \mathbf{z}_1 + \ldots + \gamma_j \mathbf{z}_j + \ldots + \gamma_a \mathbf{z}_a)}{1 + exp(\alpha_k + \gamma_1 \mathbf{z}_1 + \ldots + \gamma_j \mathbf{z}_j + \ldots + \gamma_a \mathbf{z}_a)} \quad for \;\; k = 1, \ldots, K-1 \; (1)$$

where α_k and γ_j are the intercept and slope coefficients to be estimated on the subset of a PCs. We call this approach principal component proportional odds (PCPO). We observe that any other OLR models - adjacent category, continuation ratio - could be carried out on the chosen subset of PCs.
Here, we apply the rate of well classified for measuring the performance of the model, but other criteria could be chosen (Camminatiello, Lucadamo, 2010).

3 A model to predict the Fp-ratio.

The study aims to evaluate how the quality of the marine waters depends on the continuous variables: salinity, different size of phytoplankton communities, i.e. micro-phytoplankton ($> 20\mu m$), nano-phytoplankton ($20 - 2\mu m$), and pico-phytoplankton ($< 2\mu m$)], total biomass Cholorophylla, and temperature. The Fp-ratio, utilized as the indicator of the trophic state of the Adriatic Sea, is classified in three categories: 1 ($Fp < 0.40$), 2 ($0.4 \leq Fp \leq 0.65$), 3 ($Fp > 0.65$).
The mentioned explicative variables suffer of problem of multicollinearity, in fact the Condition Index, calculated on the whole dataset, is equal to 41.77. For this reason, if we try to apply a model for OLR, we have some problems in parameter estimation. It happens both if we consider the proportionality assumption and when we do not use parallel regression hypothesis.

Our proposal to consider PCPO can be then a valid alternative to obtain the parameter estimates related to the original variables. In fact, with our method, we first consider the components Z, linear combinations of the predictors, as new explicative variables in a proportional odds model, then we select only the significant ones. In the following step, we can estimate the coefficients for the original variables, using the eigenvectors of the Principal Component Analysis. In fact, considering that

$$Z^{(a)}\gamma^{(a)} = XV^{(a)}\gamma^{(a)} = X\beta^{(a)} \quad (2)$$

where $Z^{(a)}$ are the a selected components, $\gamma^{(a)}$ are the parameter estimations related to these components, X is the matrix of original variables, $V^{(a)}$ are the eigenvectors for the selected components, we can easily obtain that:

$$\beta^{(a)} = V^{(a)}\gamma^{(a)} \tag{3}$$

Before calculating these values, it is necessary to decide which model can work better for our aims. The use of Brent test (Brent, 1990) seems not be a good idea when the explicative variables are the principal components, for this reason, we evaluate our methods considering the percentage of correct classification on a train and test (70 % and 30 %). We show the results in table 1 for Proportional Odds, Adjacent Category Model and Continuation Ratio Model both with and without proportionality assumption.

Table 1 Correct classification rate for different models

Method	% C.C. for total set	% C.C. for train set	% C.C. for test set
Proportional odds (parallel)	80.93	78.81	78.13
Proportional odds	72.09	70.20	71.88
Adjacent category (parallel)	80.47	84.11	73.43
Adjacent category	84.19	78.15	78.13
Continuation Ratio (parallel)	80.00	84.11	78.13
Continuation Ratio	84.19	78.81	78.13

It is easy to see that there are not huge differences among the methods and, in particular, the correct classification rate for the test set is equal to 78.13 % for four of six models we built. When we consider the parallel assumption we select the components 1, 3, 4 and 5, whereas, if the proportionality assumption is disgarded, we take into account components from 1 to 5.

Applying the formula 3 on the selected components and on the whole data set and considering the parallel assumption, we obtain the results synthesized in table 2.

Table 2 Coefficient estimated in function of original variables for three different models

Variables	Coefficient estimates[a]		
	Prop. odds	Adj. categ.	Contin. Ratio
Salinity	-0.348	-0.318	-0.325
Micro-Phytopl.	-0.234	-0.241	-0.244
Nano-Phytopl.	0.077	0.080	0.080
Pico-Phytopl.	0.213	0.219	0.223
Temperature	0.076	0.081	0.085
Total Biomass Chlorophyll	3.240	3.205	3.279

[a] Reference category: 1

It is easy to verify that the coefficient estimates do not differ amongs the method. There are little differences if we look at second and third decimal place, but the in-

terpretation does not change. The salinity and the presence of Micro-Phytoplankton influence negatively the probability that Fp-ratio assumes value 2 or 3 rather than 1. On the other hand, if the values of Nano-Phytoplankton, Pico-Phytoplankton, Temperature and Total Biomass Chlorophyll increase, the probability that response variable assumes values higher than one grows.

4 Remarks and perspectives

In this paper, we proposed a method for dealing with multicollinearity in OLR. The application showed that the proposed approach performs quite well on real data, hard to manage. However, an extensive simulation study is needed in order to verify that it fits well towards many situations and for selecting optimal dimension of the model.
Moreover, the rate of well classified is a performance measure widely applied in literature, but other criteria - such as the estimator variance or cross-validation - could be discussed. In the extension to other most commonly used ORL models, particular attention should be paid to the calculation of coefficients according to the original variables, when proportionality assumption is not respected. Finally, an approach for dealing with multicollinearity and outlier problems in the OLR models could be formalized.

References

1. Agresti, A.: Categorical Data Analysis. John Wiley & Sons, Inc. New York (1990).
2. Agresti, A.: Foundations of Linear and Generalized Linear Models. Wiley (2015).
3. Aguilera, A., Escabias, M.: Solving Multicollinearity in Functional Multinomial Logit Models for Nominal and Ordinal Responses. In: Dabo-Niang, S., Ferraty, F. (eds.) Functional and Operatorial Statistics. Contributions to Statistics, pp. 7-13 Physica-Verlag HD (2008)
4. Bastien, P., Esposito Vinzi, V. Tenenhaus, M.: PLS Generalised Linear Regression. Computational Statistics Data Analysis **48**, 17–46 (2005)
5. Brant, R.: Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics **46**, 1171–1178 (1990)
6. Camminatiello, I., Lucadamo, A.: Estimating multinomial logit model with multicollinear data. Asian Journal of Mathematics and Statistics **3** (2), 93–101 (2010)
7. Camminatiello, I., Lucadamo, A.: Dealing with multicollinearity and outliers in ordinal logit model. In: Bini, M., Amenta, P., D'Ambra, A., Camminatiello, I. Statistical Methods for Service Quality Evaluation, Book of short papers of IES 2019 (2019).
8. Claustre, H.: The trophic status of various oceanic provinces as revealed by phytoplankton pigment signatures. Limnology and Oceanography **39**(5), 1206–1210 (1994).
9. Hosmer, D. W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons, Inc. New York (2000)
10. Mangoni, O., Basset, A., Bergamasco, A., Carrada, G.C., Margiotta, F., Passarelli, A., Rivaro, P., Saggiomo, M., Saggiomo, V.: A case study on the application of the MSFD to mediterranean coastal systems: The po plume, as a transitional water system in the northern adriatic basin. Transitional Waters Bullettin, **7**(2), 175–201 (2013)

11. Mangoni, O., Lombardo, R., Camminatiello, I., Margiotta, F., Passarelli, A., Saggiomo, M.: Phytoplankton Community to Assess the Environmental Status of the Adriatic Sea via Non-linear Partial Least Squares Regression. Quality & Quantity, **51** (2), 799-812 (2017).
12. Nur Aidi, M., Purwaningsih, T.: Modeling Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts in Java Island. International Journal of Statistics and Applications **3**(1), 1–8 (2013)

Session of solicited contributes SS6 – *Health Quality*
Organizer and Chair: Paolo Mariani

The uneasiness index in a patient-designed quality of life questionnaire

L'indice di disagio basato su un questionario per la Qualità della Vita centrato sul il paziente

Barbara Bartolini, Serena Bertoldi, Laura Benedan, Carlotta Galeone, Paolo Mariani, Francesca Sofia, Mariangela Zenga[1]

Abstract Quality of Life questionnaires are usually designed with the main contribution of clinicians, therefore including items that are centered on the disease rather than on its multifaceted impact on people's life. In this paper, we propose an uneasiness index based on a patient-designed questionnaire (Bartolini et al., 2021) applying a pseudo-Delphi methodology combined with customer-satisfaction techniques. The uneasiness index is aimed to enhance the patients' awareness of their subjective experience with the disease and enable them to better present their situation to clinicians. The patient-designed index provides a descriptive model that can be helpful to patients, clinicians, and third parties and be further integrated with clinical details to obtain an overall view of the course of treatment for each patient.

[1] Barbara Bartolini HPS-AboutPharma srl, Milano, Italy; bbartolini@aboutpharma.com

Serena Bertoldi, Science Compass, Milano, Italy; serena.bertoldi@sciencecompass.it

Laura Benedan, Bicocca Applied Statistics Center, University of Milano-Bicocca, Milano, Italy; laura.benedan@unimib.it

Carlotta Galeone, Bicocca Applied Statistics Center, University of Milano-Bicocca, Milano, Italy; carlotta.galeone@statinfo.org

Paolo Mariani, Bicocca Applied Statistics Center, University of Milano-Bicocca, Milano, Italy; paolo.mariani@unimib.it

Francesca Sofia, Science Compass, Milano, Italy; francesca@sciencecompass.it

Mariangela Zenga, Dipartimento di Statistica e Metodi Quantitativi, University of Milano-Bicocca, Milano, Italy; mariangela.zenga@unimib.it

Abstract *I questionari sulla qualità della vita sono solitamente realizzati con un focus principalmente di tipo clinico e sono quindi incentrati sulla malattia piuttosto che sul suo impatto multi sfaccettato sulla vita delle persone. In questo articolo proponiamo un indice di disagio basato su un questionario centrato sul paziente (Bartolini et al., 2021) applicando una metodologia pseudo-Delphi combinata con tecniche di soddisfazione del cliente. L'indice di disagio serve ad aumentare la consapevolezza dei pazienti rispetto alla loro esperienza soggettiva con la malattia, allo scopo di consentire loro di presentare meglio la propria situazione ai medici. Questo indice centrato sul paziente fornisce un modello descrittivo che può essere utile a pazienti, medici e terze parti per essere ulteriormente integrato con dettagli clinici al fine di ottenere una visione complessiva del corso del trattamento di ciascun paziente.*

Keywords: Patients, Uneasiness index, Quality of Life Questionnaire

1 Introduction

Quality of life (QoL) is a broad concept that explores several aspects and functionalities of people's lives.

From a medical point of view, QoL is considered a pivotal parameter used by clinicians to evaluate how treatments and therapies influence patients' functionality and emotional state. QoL is determined by indices assessed by questionnaires that can be either generic or disease-specific (Patrick & Deyo, 1989; Rabin & de Charro, 2001; Ware et al., 2016). In general, most of the QoL questionnaires are designed with the main contribution of clinicians, therefore including items centered on the disease rather than on its multifaceted impact on people's lives. Unfortunately, a proper tool defining the patient's perception of the pathology is missing.

In this work, we propose a methodology to define a patient-designed QoL related to an uneasiness index, based predominantly on the patients' contribution.

2 The development of the QoL questionnaire

To define a patient-centric QoL tool, we used a consensus technique to favor the expression of the major players involved in dealing with the pathology. In our model, patients and healthcare professionals constitute the working group to build the settings and assertions of the questionnaire using a Pseudo-Delphi method (Boulkedid et al., 2011; Diamond et al., 2014; Marbach & Rizzi, 1991; Murphy et al., 1998; Trevelyan EG, 2015). The items of the QoL questionnaire were defined during focus groups involving a panel of patients, two clinicians, one statistician and one facilitator.

The workflow of the development of the QoL questionnaire is reported in Figure 1. The final questionnaire contains seven evaluation settings (e.g. physical, functional, emotional, family, relational, economic and medical-assistance). Within each setting, a series of assertions are generated. According to the Customer Satisfaction Techniques, every assertion is associated with a four-point Likert scale for the agreement and importance measures, so that each patient expresses the agreement and the importance of every item according to his/her own experience.

Besides, the questionnaire includes a section with structural questions exploring the current state of the disease, personal evaluation about the psychological state and the type of assistance received, and some demographic characteristics. This information completes the patient profile and can be used for further analysis and stratification.

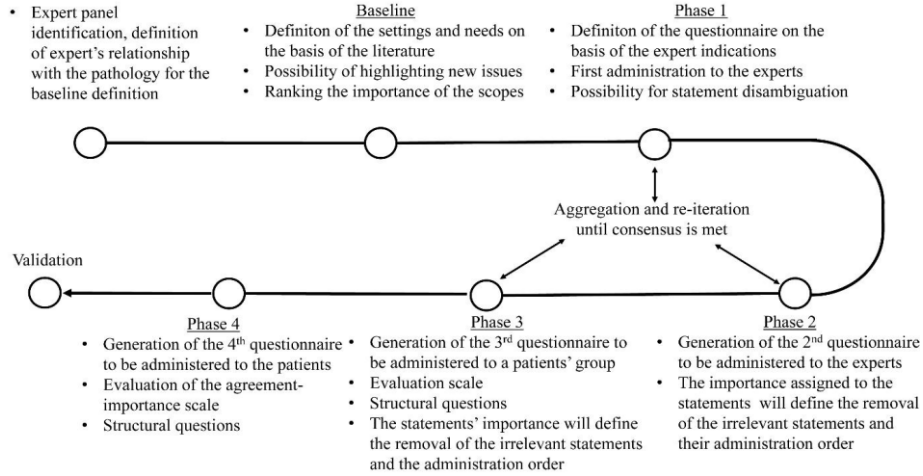


Figure 1. Flow chart showing the generation of the QoL questionnaire.

3 The uneasiness index

The methodology allows the production of a composite index for "uneasiness", which will then be compared to the internal control – provided by each patient's subjective evaluation of their own QoL on a one-to-ten scale.

Let x_{ijs} (i=1,…,n; j=1,…, k_s; s=1,...S) be the agreement of the i-*th* respondent on the j-*th* statement for the *s-th* setting. The categories on the agreement part for a statement are treated as a numeric variable: "*not at all*" =0.001; "*a little*" =0.33; "*quite a bit*" =0.67; "*very much*" =1. In this case, we transform the variable at 4 categories in 3 categories where the distance between each successive item category is equivalent and equal to 0.33. The agreement on "*not at all"* is treated as a lack pertaining to the statement. Moreover let w_{ijs} (i=1,…,n; j=1,…, k_s; s=1,...S) be the importance given by the *i-th* respondent to the *j-th* statement for the *s-th* setting. In this case the categories for the importance of a statement are "*not at all*" =0.25; "*a little*" =0.5; "*quite a bit*" =0.75; "*very much*" =1.

An indicator on the *j-th* statement for the *s-th* setting given by the *i-th* respondent is given by

$$u_{ijs} = \frac{x_{ijs}+w_{ijs}}{2} \tag{1}$$

The u_{ijs} takes values in $[0.1255;\ 1]$. For each value of u_{ijs}, it is possible to find the correct combination of $x_{ijs}\ and\ w_{ijs}$.

Table 1 displays the possible values for u_{ijs}.

Table 1: The possible values for u_{ijs} for the combinations of agreement and importance for each item.

		Importance			
		Not at all	***A little***	***Quite a bit***	***Very Much***
Agreement	**Not at all**	0.1255	0.2505	0.3755	0.5005
	A little	0.2900	0.4150	0.5400	0.6650
	Quite a bit	0.4600	0.5850	0.7100	0.8350
	Very much	0.6250	0.7500	0.8750	1.0000

For the *i-th* respondent, it is possible to create an uneasiness score for the *s-th* setting as

$$U_{is} = \sum_{j=1}^{k_s} u_{ijs} \tag{2}$$

In (2) the statements running in the opposite direction for the *s-th* setting are reversed for the score. The U_{is} could take values in $[k_s \cdot 0{,}1255;\ k_s]$.

For the *i-th* respondent, the total composite index is given by:

$$TU_i = \sum_{s=1}^{S} U_{is} \tag{3}$$

that takes values in $[0.1255 \sum_{s=1}^{S} k_s\ ;\ \sum_{s=1}^{S} k_s]$. The linear transformation of (3) in

$$TU_i^{10} = (10-1) \cdot \frac{TU_i - \sum_{s=1}^{S} k_s \cdot 0.1255}{\sum_{s=1}^{S} k_s (1-0.1255)} + 1 \tag{4}$$

allows that $TU_i^{10} \in [1;10]$. The TU_i^{10} represents the synthetic patient-centric uneasiness index. It is possible to compare TU_i^{10} with the *i-th* respondent to the score of the quality of life of the *i-th* respondent QoL_i.

3 Conclusions

In this paper, a synthetic index was proposed to evaluate the overall QoL of patients, regardless of clinical data. It enhances the patients' awareness of their subjective experience with the disease and enables them to better present their situation to clinicians.

Since this index offers a unique patient-centered perspective on their own QoL, it can provide a descriptive model helpful not only to patients, but also to clinicians and

third parties, that can be further integrated with clinical details to obtain an overall view of the course of treatment for each patient.

References

1. Bartolini, B., Bertoldi, S., Benedan, L., Galeone, C., Mariani, P., Sofia, F., Zenga, M.: Development of an innovative methodology to define patient-designed Quality of Life: a new version of a well-known concept in healthcare. In Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (eds), ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference, pp. 155-159, Firenze University Press (2021)
2. Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., Alberti, C.: Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. PLoS One **6**(6), e20476 (2011)
3. Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., Wales, P. W.: Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol **67**(4), pp. 401--409 (2014)
4. Marbach, G. Mazziotta, C., Rizzi, A.: Le previsioni. Fondamenti logici e basi statistiche). ETASLIBRI (1991)
5. Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., Marteau, T.: Consensus development methods, and their use in clinical guideline development. Health Technol Assess **2**(3), i-iv, pp. 1--88 (1998)
6. Patrick, D. L., Deyo, R. A.: Generic and disease-specific measures in assessing health status and quality of life. Med Care **27**(3 Suppl), pp. 217--232 (1989)
7. Rabin, R., de Charro, F.: EQ-5D: a measure of health status from the EuroQol Group. Ann Med **33**(5), pp. 337--343 (2001)
8. Trevelyan, E.G, Robinso, N.: Delphi methodology in health research: how to do it? Eur. J. Integr. Med. **7**(4), pp. 423--428 (2015)
9. Ware, J. E., Jr., Gandek, B., Guyer, R., Deng, N.: Standardizing disease-specific quality of life measures across multiple chronic conditions: development and initial evaluation of the QOL Disease Impact Scale (QDIS®). Health Qual Life Outcomes **14**, 84, pp. 1—16 (2016)

Assessing the Quality of Life of patients with Epidermolysis Bullosa (EB): Development of a patient-centered questionnaire

Valutazione della Qualità della Vita dei pazienti con Epidermolisi Bollosa (EB): sviluppo di un questionario centrato sul paziente

Laura Benedan, May El Hachem, Carlotta Galeone, Paolo Mariani, Cinzia Pilo, and Gianluca Tadini

Abstract Epidermolysis Bullosa (EB) is a clinical and genetic heterogeneous rare and disabling hereditary disease. EB is characterized by mucosal and skin fragility with blistering after minimal trauma. The disease and its management have a significant impact on daily life and severely affect the quality of life (QoL) of patients and their families. The present study commissioned by Fondazione REB Onlus is aimed to develop a patient-centered questionnaire to assess the QoL of EB patients using a pseudo-Delphi methodology. A multidisciplinary team including EB patients actively participated in the project. A set of domains was initially defined, and through the

Laura Benedan, Bicocca-Applied Statistics Center, University of Milano-Bicocca, Milan, Italy; email: laura.benedan@unimib.it

May El Hachem, Dermatology Unit and Genodermatosis Unit, Genetics and Rare Diseases Research Division, Bambino Gesù Children's Hospital, Rome, Italy;

Carlotta Galeone, Bicocca-Applied Statistics Center, University of Milano-Bicocca, Milan, Italy;

Paolo Mariani, Bicocca-Applied Statistics Center, University of Milano-Bicocca, Milan, Italy;

Cinzia Pilo, Fondazione REB Onlus, Milan, Italy;

Gianluca Tadini, Centro Malattie Cutanee Ereditarie, UOC Dermatologia Pediatrica ospedale Policlinico e Università degli Studi di Milano

repetition of three Delphi rounds alternated with the individual compilation of anonymous questionnaires, the specific items that compose the final version of the questionnaire were refined.

Abstract *L'Epidermolisi Bollosa (EB) è una malattia ereditaria rara e invalidante, clinicamente e geneticamente eterogenea. È caratterizzata da fragilità delle mucose e della pelle con formazione di vesciche che possono manifestarsi a causa della più lieve frizione. La malattia e la sua gestione hanno un impatto significativo sulla vita quotidiana e influenzano gravemente la qualità della vita (QoL) dei pazienti e delle loro famiglie. Il presente studio, commissionato da Fondazione REB Onlus, ha l'obiettivo di sviluppare un questionario centrato sul paziente in grado di valutare la QoL dei pazienti EB utilizzando la metodologia pseudo-Delphi. Un gruppo multidisciplinare comprendente pazienti EB ha partecipato attivamente al progetto. È stata inizialmente definita una serie di ambiti e attraverso la ripetizione di tre tavoli di lavoro Delphi alternati alla compilazione individuale di questionari anonimi è stato effettuato l'affinamento degli item specifici che compongono la versione finale del questionario.*

Key words: Epidermolysis Bullosa, Patient-centered approach, Quality of Life, Pseudo-Delphi methodology.

1 Introduction

Epidermolysis Bullosa (EB) is a group of rare genetic disorders that are clinically heterogeneous and encompass a broad spectrum of severity. Most EB subtypes share extreme fragility of the skin and mucous membranes as the key symptom, which leads to a tendency to develop blisters and skin wounds after mild trauma.

The disease may have several mucocutaneous and systemic manifestations, including blisters in the oral mucosa and esophagus, dental problems, dysphagia, hair loss, severe anemia, malnutrition, osteoporosis, pseudosyndactyly, contractures, renal failure, cardiopathy. Everyday activities represent a constant challenge because of disease-associated functional limitations and the frequent need for painful and time-consuming medications. EB patients' unmet needs go beyond requests for medical support (Dures, Morris, Gleeson, & Rumsey, 2011). Indeed, esthetical and functional damages lead to additional psychological burden and impact on social relationships. Lack of awareness and understanding by both laypeople and non-specialist healthcare professionals represent a further issue. As a result, the Quality of Life (QoL) of patients (and their families) is severely compromised.

A valid and reliable tool to assess the QoL of these patients may offer valuable insight to better understand each patient's experience and to improve patient care and management. This topic has already been tackled by the international literature: the QoLEB questionnaire (Frew, Martin, Nijsten, & Murrell, 2009) emerged as the most used instrument available to assess EB patients' QoL. It was initially developed and validated in English with an Australian sample, resulting reliable. To our knowledge,

all existing EB QoL questionnaires have been developed by clinicians, and there is no tool developed by patients to investigate the QoL in EB from the patient's perspective yet. This questionnaire aimed to investigate the quality of life on a broad spectrum by considering aspects relating to the psychological, social, personal sphere beyond the mere clinical symptoms. In our opinion, a patient-centered approach was deemed to be the most impactful for a deeper understanding of patients' perspective, offering them the opportunity to make their voices heard. The Delphi methodology was chosen as the most appropriate way to reach our goal.

2 Methodology

The project was launched by the Italian non-profit Foundation for EB Research and the Italian EB Registry, Fondazione REB Onlus. They highlighted the need to develop a patient-centered questionnaire to assess the QoL of adult patients with EB, to provide a valid tool for clinicians to identify the areas that need more attention in order to improve global patient care.
Firstly, a critical literature review was conducted to understand the topic from a theoretical perspective. Hence, an online Delphi study was organized to grasp the first-hand experience of EB patients.

2.1 Pseudo-Delphi Method

The Delphi method is a flexible and iterative process known to be a valuable method to be applied in health research (Trevelyan & Robinson, 2015). It consists of envisaging some key topics to a specific panel of experts and subsequently providing evaluations through repetition and reassessing a previously discussed solution to reach a consensus, as the final expression of the group opinion (Marbach, Mazziotta, & Rizzi, 1991). This method was deemed appropriate for analyzing real-world data pertaining to a highly subjective matter - such as QoL – from the perspective of rare disease patients.

2.2 Study procedure

A multidisciplinary panel was recruited comprising: a Delphi master with a solid statistical and methodological expertise; a group moderator; six patients or child patients' caregivers; two clinicians recognized as international key opinion leaders on EB; a psychologist.

In this study, the Delphi procedure should be considered "Pseudo-Delphi" because all the group discussions were organized via "face-to-face" virtual meetings. Each participant had the opportunity to share his/her opinions and contribute to the group

discussion. Nonetheless, after each group meeting, a private evaluation of the discussed topics was granted, having anonymity assured, allowing each expert to analyze and re-consider all items composing the questionnaire critically, make suggestions, express comments and provide individual responses without any social pressure or compliance effect that may conversely arise during the group discussions. All the answers were then summarized to be presented to the group as an aggregate and anonymous contribution in order to reach a consensus. The overall study procedure is outlined and displayed in Figure 1.

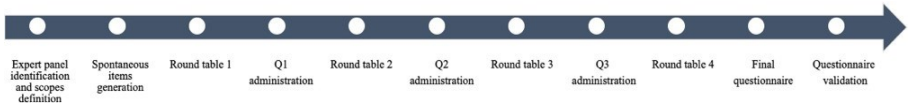


Figure 1: *Study procedure*

All experts were invited to a kick-off meeting to present the project, illustrate the methodology and discuss the study aims. Successively, the patients and clinicians were asked to work individually and create a list of spontaneously generated items addressing all relevant aspects of their QoL. All the answers were collected anonymously. More than 160 sentences were registered, and they were carefully considered, analyzed and grouped within specific domains. Items with the same meaning were combined, duplicates were discarded, and overall harmonization of all the statements was carried out.

The results were presented in the first round table, during which the experts contributed to the discussion openly. After that, the first version of the questionnaire (Q1) was created and sent to each participant individually for anonymous completion. Each person was asked to read every statement, assess their degree of importance, comment on each item's clarity and specificity, and add anything relevant that might have been omitted. Considering the degree of importance indicated by the participants, a ranking was created among the items within every domain. The results of this analysis were further discussed during the second round table to make it possible to refine the questionnaire removing the items of lower importance.

The second version of the questionnaire (Q2) was administered to the experts who filled it in anonymously. This time, participants were required to rate both the degree of agreement and the degree of importance of each item on a four-point Likert scale ("Not at all = 1", "A little = 2", "Quite = 3", "Very = 4").

The results of this phase were discussed during the third round table. Further refinement of the items grouped in the abovementioned seven domains and the specific questions was made.

After that, the new version (Q3) was prepared and administered to the experts, who anonymously filled it in. All experts received an 85-item questionnaire assessing EB patients' QoL examining seven different life domains:

- *Physical* comprises the most relevant aspects in terms of health and physical well-being.

- *Functional ability and autonomy* includes statements about self-sufficiency and the ability to perform daily and routine actions.
- *Psychological-emotional* refers to sensations, emotions, thoughts and feelings that may affect the psycho-emotional well-being.
- *Family* includes statements concerning relationships with parents, brothers and sisters, or other family members, possibly including partners and children.
- *Relational* addresses relationships and frequent interactions with people who do not belong to the family (e.g., friends, classmates, colleagues, strangers on the street, etc.).
- *Work and economic* relates to the work context and the financial implications of the disease.
- *Medical care and assistance* includes aspects about disease-related healthcare, such as medical and nursing assistance (see Benedan et al., 2021 for more detailed information about the study procedure and questionnaires composition).

3 Results and Discussion

Six patients returned Q3, and all their answers were analyzed according to the Customer Satisfaction techniques.

Firstly, all the answers to items with a negative meaning were reversed (e.g., *"I feel like I'm a burden on my family"*) so that all the scores could be interpreted in the same direction: the higher the score, the better the QoL.

The scores within each domain were summed up for every person. All the values about each item's degree of agreement and importance were combined, and the average score was computed. The mean values of "Agreement", "Importance", and their combined score for each domain are displayed in Table 1.

Table 1: Mean values for agreement and importance of each domain (N=6)

Domain	***Agreement***	***Importance***	***Mean score Agreement-Importance***
Physical	2.13	1.54	1.84
Functional ability and autonomy	2.22	2.75	2.49
Psychological-emotional	2.75	2.32	2.54
Family	2.55	1.81	2.18
Relational	2.63	2.13	2.38
Work and economic	2.56	2.49	2.53
Medical care and assistance	2.25	2.50	2.38

The results showed differences among domains. The physical domain represents the most compromised area of QoL because of the relevance and severity of clinical

features and symptoms. Nonetheless, other domains may be seen as a valuable resource when considering the quality of life of people affected by a disabling disease such as EB. Three domains, in particular, seem to be important in improving the patients' QoL and deserve further attention in future studies: psychological and emotional well-being; capability to manage functional autonomy in everyday life; ability to work despite the disease-related limitations.

4 Conclusions

The present study is part of a more extensive research project aimed at developing a valid and reliable patient-centered questionnaire to assess the QoL of adult EB patients. The results showed that the concept of QoL should consider the patient's overall experience beyond clinical classifications and not limited to physical symptoms. The patient-centered questionnaire here evaluated comprised 85 items grouped in seven domains encompassing physical manifestations, functional autonomy, psycho-emotional state, social relations inside and outside the family context, the economic and working field and several aspects of medical care and assistance. Six patients completed the questionnaire assessing their level of agreement and the importance of every statement. The results highlighted differences among the domains in determining the individual and subjective patient's QoL.
The future steps of this research will be addressed to assess the questionnaire's psychometric properties to prove its reliability in measuring the QoL in a larger sample of adult EB patients. Although this study has been conducted in Italy, it may be translated, trans-culturally interpreted and validated to become a universally valid tool to assess the whole range of QoL domains affected by this disease.

References

1. Benedan, L., El Hachem, M., Galeone, C., Mariani, P., Pilo, C., Tadini, G.: Patient-generated evidence in Epidermolysis Bullosa (EB): Development of a questionnaire to assess the Quality of Life. ASA 2021 Statistics and Information Systems for Policy Evaluations. In: Bertaccini, B., Fabbris, L., Petrucci, A. (eds) Book of short papers of the on-site conference (2021).
2. Dures, E., Morris, M., Gleeson, K., Rumsey, N: The psychosocial impact of epidermolysis bullosa. Qual. Health Res. **21**, 771–782 (2011).
3. Frew, J.W., Martin, L.K., Nijsten, T., Murrell, D.F.: Quality of life evaluation in epidermolysis bullosa (EB) through the development of the QOLEB questionnaire: An EB- specific quality of life instrument. Br. J. Dermatol. **161**, 1323–1330 (2009).
4. Marbach, G., Mazziotta, C., Rizzi, A.: Le previsioni. Fondamenti logici e basi statistiche. Milano: edizioni Etaslibri (1991).
5. Trevelyan, E. G., Robinson, N.L.: Delphi methodology in health research: How to do it? European J. Integr. Med. **7**, 423–428 (2015)

Survival analysis in a business context: how to control the abandons of my subscribers

Analisi di sopravvivenza in contesto aziendale: come controllare gli abbandoni fra i miei clienti

Andrea Marletta and Marco Morandi

Abstract The statistical literature proposed many contributions about survival analysis in medical research, in this work this approach is proposed in a business context. The aim of this paper is to control the mortality of the users belonging to an e-mail subscribers list for a company operating in the healthcare information sector. Having available the survival times for each subscriber, the choice was oriented to survival models to evaluate the abandon of the customers. A survival analysis was conducted through a Cox model considering some risk factors of the subscriber. The selected Cox model carried to the identification of risk profiles representing different situations in terms of probability of abandon.

Abstract *La letteratura statistica ha proposto molti lavori riguardanti l'analisi di sopravvivenza nella ricerca medica, in questo lavoro questo approccio è proposto in un contesto aziendale. Lo scopo del contributo è modellare la mortalità degli utenti appartenenti ad una lista di iscritti via mail per una azienda operante nel settore dell'informazione medico-scientifica. Avendo a disposizione i tempi di sopravvivenza per ogni iscritto, la scelta è stata orientata su modelli di sopravvivenza per valutare gli abbandoni dei clienti. L'analisi di sopravvivenza è stata effettuata attraverso un modello di Cox considerando alcuni fattori di rischio per i clienti. Il modello finale selezionato ha portato all'identificazione di categorie di clienti più a rischio rappresentando situazioni differenti in termini di probabilità di abbandono.*

Key words: Survival analysis, e-mail marketing, Cox model

Andrea Marletta
University of Milano-Bicocca e-mail: andrea.marletta@unimib.it

Marco Morandi
PKE e-mail: m.morandi@pke.it

1 Introduction

During last years, among the marketing strategies, one of the most used is the e-mail marketing. Companies are using email marketing to engage with customers and encourage active transactional behavior. Extant research either focuses only on how customers respond to email messages or looks at the average effect of email on transactional behavior [7].

The analysis was based on research proposed by PKE, Professional Knowledge Empowerment, a company created to manage Italian healthcare databases. Over time, the areas of expertise have expanded, thus specialising both in data management and communication. In the communication area, one of the services is the e-mail marketing. From an increasingly digital standpoint, communication strategies must also take into account the change that PKE reinterprets, by making email marketing projects available that guarantee precious and exclusive value: in-depth knowledge of the health professional and in particular of doctors.

In this paper, this strategy is faced from a statistical point of view. Other authors tried to deal with this issue using a Bayesian approach [1, 6], here the used technique is the survival analysis. To apply this method, it is necessary to have a time variable measuring the difference between the birth and the death of the phenomenon. About the e-mail marketing, the birth time could be represented by the entrance of the customer in the subscriber list and the death the exit from this list.

Starting from this analogy between the concept of death in a natural population and the e-mail marketing, the idea is to use some statistical techniques usually used for survival analysis as models able to predict information on the subscribers. Following this approach, parametric (Weibull) and semi-parametric models (Cox) have been applied for the available data. The aim of this paper is to create an useful tool to follow the temporal evolution of the profiles. This could be done creating different strategies on the basis of the features of the customer.

The paper is structured as follows: after the introduction, a second section is dedicated to the methodologies used to answer the research objectives. A third section will show the description of the dataset and some preliminary results.

2 Survival analysis

Survival analysis contains all the techniques and statistical models designed for the description and the analysis of time events of a statistical unit. It is necessary to identify the unit exposed at risk respect to this event and the measure of the time duration and the end of these event.

Survival is therefore characterised by a time variable with a start-up and an end-point. In medical research, start-up corresponds to time in which an individual has been introduced in the experimental study or a clinical treatment or the start of a particular condition for a disease. On the other hand, if the end-point is the death of the

patient, data are referred to the death time. The end-point could be not necessarily the death, but also the end of a pathological state.

For this work, the start-up is the date in which the customer was subscribed in the e-mail lists and the end-point is represented by the exit of the customers from the list.

Survival data own some features that need the use of some tailored statistical procedures. The first one is their distribution, generally survival data are not symmetrically distributed: an histogram based on survival times will tend to be positively asymmetric; this means that all classical models as linear regression are not suitable for these data. The second one is the presence of censored data, that is to say, statistical observation that did not experiment the time event. In medical research, they are all patient are not dead at the end of the experiment or dead for alternative causes or retired from the treatment.

Survival analysis can be treated using non-parametric, parametric or semi-parametric models. The first non-parametric approach considers the estimate of the survival function of a t time variable using the life-tables. These tables are obtained dividing the observation period in temporal intervals [2].

Non-parametric models are very flexible but they do not guarantee consistent and precise estimates. This is why they are usually as exploratory tools. For this reason, parametric models have been proposed proposing that the time variable assumes a probability distribution depending on some parameters. This approach allows to determine possible combinations of explanatory variables or risk factors conditioning the risk and the survival function. Once the probability distribution function $(f(t))$ is chosen, then it is possible to obtain the survival function $(S(t))$, the hazard risk function $(h(t))$ and the cumulative hazard risk function $(H(t))$. The most used probability function for time variables are the exponential and the Weibull distribution.

Finally, semi-parametric models were introduced by Cox [3] in 1972 and it is so defined because even if it is based on the hypothesis of proportional hazards, it makes no assumption about a probability distribution for the survival times. The Cox model assumes the hazard risk function $h_i(t)$ as a product of two components:

$$h_i(t) = h_0(t) * exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \tag{1}$$

The first component $h_0(t)$ is named baseline hazard risk function, the second one is the exponential of the sum of the combination terms $\beta_i x_i$ extended to all p explanatory variables.

Survival analysis is principally used in medical statistics, but there are a lot of applications of this method in economic issues. The application of survival analysis in economic field could have companies or obligors as statistical units.

Giambona and Vassallo [4] in 2007 built hazard risk profiles for Italian banking credits using a survival model at discrete time on loan data from Italian banking system. They want to observe the hazard risk function in the first decade after the loan assignment. The hazard risk profiles were built using the risk levels compared to the baseline profile through odds ratios. Giambona also studied the death rate of Italian banking credits using a non-proportional hazard logistic model [5].

3 Application

In this paper, the dataset was available thanks to PKE and it is composed by all the subscribers in their e-mail marketing list. PKE sends over 18 million emails every month, this tool has allowed it to perfect communication models for promotion of drugs in launch, mature or in decline, in siding or replacing the local pharmaceutical representative. The target audience is made of pharmaceutical companies, medical device companies, certification bodies, scientific societies, patient associations, insurance, technology companies, public/private bodies of the NHS, CME providers, publishing companies, public utilities.

Several models could be obtained considering different dependent variables of the Cox model. A time variable could be computed as difference between the subscription in the list and the last received e-mail. Another time variable could be the difference between the subscription and the last time the subscriber opened or clicked the e-mail.

The risk factors included in the Cox model are the number of received mails and the feature of the subscriber. The available information are gender, age, workplace, the dummy variable about activity profile (1 for active, 0 for non-active), the belonging to a category of the target audience and their specialization.

Once these Cox models are estimated, it is possible to define some risk profiles and determine the categories of target audience more inclined to abandon the e-mail marketing strategy.

References

1. Ansari, A., Mela, C. F. E-customization. Journal of marketing research **40(2)**, 131-145, (2003).
2. Collett, D. Modelling Survival Data in Medical Research. Chapman & Hall: London, (1994).
3. Cox, D.R. Regression models and life-tables (with discussion). Journal of Royal Statistical Society, Series B **74**: 187–220, (1972).
4. Giambona, F., Vassallo E. Profili di rischio dei crediti bancari italiani: un'analisi per generazioni di finanziamenti, Rivista Minerva Bancaria **2**, 9-46, (2007).
5. Giambona, F. Mortalità dei crediti bancari italiani: Altre evidenze empiriche, Rivista Minerva Bancaria **5**, 1-16, (2007).
6. Wu, J., Li, K. J., Liu, J. S. Bayesian inference for assessing effects of email marketing campaigns. Journal of Business & Economic Statistics **36(2)**, 253-266, (2018).
7. Zhang, X. Managing a Profitable Interactive Email Marketing Program: Modeling and Analysis. Georgia State University, (2015).

Session of free contributes SCL5 – *Assessing Performance*
Chair: Cristina Davino

Political performance measuring and tracking through a system based on the Political Performance Indicator (I_{ep}): Naples 2021 case

Monitoraggio e misurazione dell'efficienza politica attraverso un sistema basato sull'indicatore di efficienza politica (I_{ep}): il caso Napoli 2021

Giovanni Di Trapani

Abstract Many studies have been carried out on election systems; in fact, literature on this matter is very broad ranging and multidisciplinary, spanning from law to statistics. This paper examines the electoral administrative systems, and in trying to focus on the political supply analysis, concentrates its attention to the pre-electoral process, the presentation of electoral groups and the electorate. Using a multi-disciplinary approach, this paper brings current regulations and quantitative methodology of analysis into line with the aim of achieving a clear understanding of the administrative vote and the political offer. The aim of the paper is the measurement and monitoring of the efficiency of political systems through an index of organisational efficiency (I_{eo}) for political parties. At the beginning, the processes identified by current regulations as crucial during the so-called pre-electoral phases of preparation of political parties' groups (Signature presentation and Candidates) were identified in two. The first refers to the signing of the declaration by the list presenters, the second to the declarations of candidate acceptances. The main purpose of this paper is to suggest a synthetic index comprising several indicators designed to measure the efficiency (and observance) of current legislative provisions on elections

[1] Giovanni Di Trapani, researcher at CNR - IRISS Via Card. G. Sanfelice, 8 – 80100 Napoli (I)– E-Mail: g.ditrapani@iriss.cnr.it

Giovanni Di Trapani

Abstract *Numerosi sono gli studi condotti sui sistemi elettorali; la letteratura sull'argomento è, infatti, molto ampia e variegata e certamente multidisciplinare, spaziando dal diritto alla statistica. Il presente lavoro approccia i sistemi elettorali amministrativi, e nel tentativo di contribuire ad un focus sull'analisi dell'offerta politica, focalizza l'attenzione sulla fase pre-elettorale, quella della presentazione delle liste elettorali e dell'elettorato passivo. Con un approccio multidisciplinare, il presente lavoro coniuga la normativa vigente con l'impiego di una metodologia d'analisi statistico quantitativa ed intende addivenire ad un focus sul voto amministrativo e dell'offerta politica.*

Lo scopo del lavoro è la misura ed il monitoraggio dell'efficienza dei sistemi politici attraverso la della costruzione di un indice dell'efficienza organizzativa (I_{eo}) delle forze politiche. Inizialmente, sono stati identificati in due i processi ritenuti dalla normativa vigente come fondamentali nella fase c.d. pre-elettorale della preparazione delle liste (Presentatori delle firme e Candidature). Il primo processo fa riferimento alla fase di sottoscrizione della dichiarazione da parte dei presentatori delle liste, il secondo alle dichiarazioni di accettazione delle candidature. Obiettivo primario del lavoro è quello di proporre un indicatore sintetico espressione di alcuni indici con il precipuo compito di fornire una misura dell'efficacia (e del rispetto) delle disposizioni normative vigenti in materia elettorale

Key words: Electoral statistics, Political marketing, Efficiency of organization in lists and coalitions.

1 Introduction

Many studies have been carried out on election systems; the literature on the topic is, in fact, extensive, varied and certainly multidisciplinary, ranging from law to statistics. The theme of election systems is hot and wide opened (Ceccanti, 2017) because " any election system is a real and proper filter of society and politics" (Pasquino, 1984). In analyzing an election system, various elements need to be considered singly and overall, in the interrelations between the different components and among those elements to consider aspects, at times, not easily perceived.

An electoral system is a complex of rules serving to discipline the participation of citizens and the ways of representation, but also a way of influencing party formation and the recruitment of political leaders. In Italy, the political instruments allowing citizens entitled to vote to express a personal preference are distinguished as political and administrative. The first ones allow the election of House and Senate members, while with the administrative type of local voters could choose their local administrators, such as mayors and councilors.

This paper examines administrative electoral systems, and to contribute a focus to the analysis of the supply of politics, concentrates the attention to the pre-electoral step, namely the presentation of voters' lists and the voter. Using a multidisciplinary approach, the paper joins current regulations to the use of a statistical-quantitative approach and intends to reach a focused study of the administrative election and of the supply of politics. Traditional analysis of political supply is concentrated merely on socio-political aspects, focusing primarily on measures of forces in play and concentrating on aspects related to coalition composition - seen as combinations of candidate groups - supporting different candidates for Mayor. An approach which provides interesting insights on the quality composition of the parties but limited to the age of the candidates or, at most, the gender gap in the electorate. The paper, then, by presenting a dashboard of indicators and an index, aims to propose a focus on the supply of politics measuring the " organizational efficiency " of the opposing political parties (coalitions and lists) participating in the elections of municipalities. The procedures and operations concerning the submission and admission of candidatures for the direct election of the mayor and the municipal council in the cities of the ordinary statute regions are certainly complex; for the purpose of providing the responsible organs a unique guideline in this delicate phase, the Dipartimento per gli affari interni e territoriali and more precisely the Direzione centrale per i servizi elettorali annually provides a vademecum for administrative elections. These ministerial publications contain all the bureaucratic formalities for the presentation of electoral lists.

1.1 Methodology

To construct the index of organizational efficiency (I_{eo}) for political parties, initially two processes have been identified which are considered fundamental by current regulations during the pre-electoral stage, namely, the process of the list presenters and the corresponding candidatures. The first process relates to the phase of subscription of the Declaration by list submitters, and the second to statements of acceptance of candidatures.

The focus of this paper is to suggest some synthetic indicators whose primary purpose is to provide a measure of the efficacy (and compliance) with current electoral regulations. In the first process, the indicator of subscriptions (I_s) was identified, covering the limits of signatures required for the submission of the lists taking into consideration possible exceptions for parties already represented in the two Chambers of the Italian Parliament. Referring, instead, to the second process - more complex - have been considered the number of the candidates, by identifying an indicator of candidacy (I_c) and of an indicator of the proportion of gender representation (I_{rg}). The model proposes, finally, a further indicator related to

political complexity (Icp), that identifies and measures the quantitative composition of coalitions, understood as an aggregation of lists of candidates supporting mayoral candidates.

Table 1: Indicators

Indicator description	***Abbreviation***
Subscription's indicator	Is
Candidate indicator	I_c
Gender representation ratio indicator	I_{rg}
Political complexity indicator	I_{cp}

1.1.1 Subscriptions' indicator (I_s)

The declarations of presentation of the lists of candidates for the municipal council and the associated candidacies for the office of mayor, for each municipality, must be signed - in accordance with article 3, paragraph 1, of Law no. 81 of March 25, 1993, as amended, as well as article 2, paragraph 1, of Decree-Law no. 25 of March 5, 2021, no. 25, converted, with amendments, by Law no. 58 of May 3, 2021 (which, for the year 2021, reduced the minimum number of subscribers to one third) - by a certain number of voters in the municipality, depending on the relevant population bracket. The indicator Is (Equation 1) aims to represent this first condition provided by the legislation and is given by the average value per coalition of the incremental ratio between the number of list signatures (subscriptions) and the respective maximum number.

Equation 1

$$\sum_{i=1}^{n} \frac{(X_n - X_{min})/(X_{max} - X_{min})}{n}$$

the indicator takes the following values: 0 for the number of list signatures equal to the minimum number provided and 1 for the number of list signatures equaling the maximum number provided.

1.1.2 Candidate indicator (I_c)

Each candidate's list should also be accompanied by a declaration of acceptance of candidature on the part of each candidate for the office of mayor or city councilor, pursuant to and in accordance with article 28, paragraph four, and article 32, paragraph seven, number 2), of the Consolidation Act no. 570/1960. Each list must include several candidates not exceeding the number of councilmen to be elected in the council and not less than two thirds in compliance with Article 73, paragraph 1,

and Article 37, paragraph 1, of Legislative Decree no. 267/2000. When the number of municipal councilors to be elected is not exactly divisible by 3, for the determination of the minimum number the above-mentioned article 73, paragraph 1, is applied, based on which, when the number of candidates to be included in each list, resulting from the aforesaid calculation, contains a decimal figure greater than 50, it is rounded up to the next higher unit.
Similarly, to the calculation of the I_s indicator, the Candidate Indicator (I_C) is calculated as the average value per coalition of the incremental ratio between the number of candidates on the list and the respective maximum number.

Equation 2

$$\sum_{i=1}^{n} \frac{(X_n - X_{min})/(X_{max} - X_{min})}{n}$$

This will assume values equal to 0 for the number of list members equal to the minimum number envisaged and equal to 1 for the number of list members equal to the maximum number envisaged.

1.1.3 Gender representation ratio indicator (I_{rg})

Pursuant to Law No. 215 of November 23, 2012, in Municipalities with a population of more than 15,000 inhabitants, the lists of candidates must be formed in such a manner that each gender is not represented by less than one-third and no more than two-thirds of the number of candidates. It is determined by the minimum value of I_m masculinity and I_f femininity indices per coalition

Equation 3 I_m e I_f

$$\sum_{i=1}^{n} \frac{m_n}{n_i} \qquad\qquad \sum_{i=1}^{n} \frac{f_n}{n_i}$$

The indicator of the proportion of gender representation (I_{rg}) is, therefore, calculated according to the following equation

Equation 4

$$min\ (I_m\ ;\ I_f\ )$$

1.1.4 Political complexity indicator (I_{cp})

In Municipalities with a population of over 15,000 inhabitants, each candidate for the office of mayor must declare - in addition to acceptance of the candidacy and the absence of the condition of ineligibility - the connection with the list or lists presented for the election of the municipal council. For this reason, the average value of the ratio between the number of lists per coalition (X_c) and the total number of lists (X_C) has been calculated in Equation 5.

Equation 5

$$\sum_{i=1}^{n} \frac{(X_c)/(X_C)}{n}$$

The purpose of the work is, therefore, the construction of an index intended as a mathematical association of the set of proposed indicators (variables) that represent the different components of the system for measuring the political supply. This index assumes a character proportional to the previous indicators- and is expressed by the following equation:

Equation 6

$$\frac{I_S \times I_C \times I_{Rg} \times I_{Cp}}{n}$$

The paper pursues a non-compensatory approach and employs a model for measuring indicators that are formative in nature, and therefore not interchangeable, and whose correlations are not explained by the model itself. To validate the proposed synthetic index, with the aim of verifying that the index is consistent with the general theoretical framework, the work aimed to test the indicators and the index of organizational efficiency (I_{eo}) through an application of the model to the election round of the administrative 2021 and precisely to the elections of the municipality of Naples. Finally, with the specific purpose of comparing the variables that make up the proposed model, the values generated by the contextualization to the so-called case Napoli'21 are also compared and contrasted with the previous two rounds of elections in the same territorial context and precisely those of June 5, 2016 and May 2011.

2 References

1. Angelucci, D., Paparo, A.: Comunali: equilibrio, stabilità e il ritorno del bipolarismo, Centro Italiano Studi Elettorali, Permanent link: https://cise.luiss.it/cise/cise2019_wp/2019/06/13/comunaliequilibrio-stabilita-e-il-ritorno-del-bipolarismo/ (2019)
2. Botti, G.: I risultati ei flussi elettorali nella cintura di Napoli - cise.luiss.it http://cise.luiss.it/cise/wp-content/uploads/2017/08/DCISE9_2-16.pdf (2017)

Political performance measuring and tracking through a system based on the Political Performance Indicator (PIE): Naples 2021 case Contribution

3. Brancaccio, L., Dines, N.; Pine, J., Ravveduto, M.: Dentro la città, Meridiana: rivista di storia e scienze sociali. **80**, 2, ISSN: 1973-2244 P. 197-220 DOI: 10.1400/223978 Permalink: http://digital.casalini.it/10.1400/223978 (2014)
4. Brancaccio, L., Zaccaria, A.M.: Verso la città dei municipi: la dimensione territoriale della politica a Napoli Liguori Editore Srl, 177 pagine (2007)
5. Brancaccio, L.: Crisi del clientelismo di partito e piccole rappresentanze territoriali. Forme e spazi del consenso personale a Napoli Quaderni di Sociologia N. **78** Pagine 77-99 DOI: https://doi.org/10.4000/qds.2169 Permalink: https://journals.openedition.org/qds/2169 (2018)
6. Ceccanti, S.: I sistemi elettorali nella storia della Repubblica: dalla Costituente alla legge Rosato. Federalismi.It, **20**, 1–9 (2017)
7. Chiaramonte, A., E.manuele, V.: Multipolarismo a geometria variabile: il sistema partitico delle città, in Emanuele V., Maggini N., Paparo A. (a cura di): Cosa succede in città? Le elezioni comunali 2016, Dossier CISE (**8**), Roma, Centro Italiano Studi Elettorali, pp. 129-137 (2016)
8. Di Trapani, G.: Focus sugli indicatori della dinamica migratoria. Paradox, **4-5** (Maggio-Giugno 2016), 1-6. http://www.rivistaparadox.it/archivio/rivista-n-4-5-luglio-novembre/item/88-focus-sugli-indicatori-della-dinamica-migratoria.html (2016)
9. Di Trapani, G.: Rapporto macroeconomico del divario Nord-Sud in Italia. Paradox, **2** (Marzo-Aprile 2016). http://www.rivistaparadox.it/archivio/rivista-n-2-marzo-aprile/item/33-rapporto-macroeconomico-del-divario-nord-sud-in-italia.html (2016)
10. Di Trapani, G.: È allarme: l'analisi degli indicatori della dinamica demografica. Paradox, **3** (Maggio-Giugno 2016). http://www.rivistaparadox.it/archivio/rivista-n-3-maggio-giugno/item/59-e-allarme-l-analisi-degli-indicatori-della-dinamica-demografica.html (2016)
11. Emanuele, V.: Riscoprire il territorio: dimensione demografica dei comuni e comportamento elettorale in Italia, Meridiana, **70**, pp. 115-148 (2011)
12. Emanuele, Vincenzo; Maggini, Nicola: Cosa succede in città? Le elezioni comunali 2016 Edition: Dossier CISE **8** Publisher: CISEISBN: 978-88-98012-19-0 Permalink: https://www.researchgate.net/publication/305073459_Cosa_succede_in_citta_Le_elezioni_comunali_2016 (2016)
13. Goodman, L. A.: Ecological regressions and behavior of individuals. American Sociological Review, 18, 663-664. https://doi.org/10.2307/2088121 (1953)
14. Martone, V.; Brancaccio, L.: Nuove strategie di consenso a Napoli: il ceto politico nel decentramento comunale, Meridiana: rivista di storia e scienze sociali: **70**, 1, ISSN: 1973-2244 P. 17-48 DOI: 10.1400/183035 Permalink: http://digital.casalini.it/10.1400/183035 (2011)
15. Ministero Interno: Istruzioni per la presentazione e l'ammissione delle candidature - Elezione diretta del sindaco e del consiglio comunale a cura del Dipartimento per gli affari interni e territoriali - Direzione centrale per i servizi elettorali; pubbl. n. **01** Consiglio Comunale (amministrative); Agosto'21 (2021)
16. Pasquino, G.: Rappresentanza politica, sistema elettorale e formazione del Governo: una proposta. Il Mulino, **4**, 660–673. https://doi.org/DOI: 10.1402/14046 (1984).
17. Prete M. Irene: Aspetti metodologici e strategici dell'approccio di marketing politico Università del Salento – Coordinamento SIBA eISBN 978-88-8305-113-5 (electronic version) DOI Code: 10.1285/i9788883051135 http://siba-ese.unisalento.it (2015)
18. Tufte, E.R.: Determinants of the Outcomes of Midterm Congressional Elections, American Political Science Review **69**(3), pp. 812-826 (1975)

A class of case-mix adjusted probability-based indices for performance evaluation

Una classe di indici di performance basati sulla probabilità e aggiustati per il case-mix

Giorgio E. Montanari and Marco Doretti

Abstract A family of indices for the assessment of the performance of a group of agencies providing a certain service is introduced to deal with settings where binary variables are used as outcome measures. The indices are built by aggregating group-specific success probability contrasts in order to account for case-mix, that is, users' characteristics influencing the outcome and thereby acting as confounders. The proposed family encompasses many particular cases allowing to weight differently performance improvements or worsenings as well as performance shifts occurring at distinct average levels. Therefore, it represents a flexible tool adapting to various applied evaluation contexts. A general framework for estimation is also discussed.
Abstract *In questo lavoro viene introdotta una famiglia di indici per la valutazione della performance di un gruppo di agenzie erogatrici di un servizio, da impiegare in presenza di variabili risultato binarie. Gli indici sono costruiti aggregando contrasti tra probabilità di successo relative a distinti gruppi di utenza, opportunamente definiti per considerare il case-mix ed eliminare effetti di confondimento. La famiglia proposta comprende molti casi particolari che consentono di ponderare in maniera diversa miglioramenti o peggioramenti di performance, così come discrepanze rispetto a livelli medi differenti. Per questo motivo, essa rappresenta uno strumento di valutazione in grado di adattarsi a svariati contesti applicativi. Infine, viene presentato un approccio generale per la stima degli indici.*

Key words: binary outcome, case-mix, generalized performance index, quality assessment.

Giorgio E. Montanari
University of Perugia, Department of Political Science, via A. Pascoli 20, 06123 Perugia (Italy), e-mail: giorgio@montanari.unipg.it

Marco Doretti
University of Perugia, Department of Political Science, via A. Pascoli 20, 06123 Perugia (Italy) e-mail: marco@doretti.unipg.it

1 Introduction

Service monitoring has nowadays become a standard process in many areas of the public as well as of the private sector. In this regard, efficacy remains a crucial dimension of the overall quality level, together with efficiency and equity/accessibility [3]. Quite often, the problem arises of measuring the efficacy of a number of supplying agencies by means of observational data recorded on their users. As typical in these non-experimental settings, the outcome variables employed in this evaluation process are influenced by both agencies' abilities and users' personal characteristics acting as confounders. The latter form the so-called *case-mix*, and have to be accounted for in order to draw correct conclusions and make fair comparisons among agencies.

In some contexts, classification systems have been introduced which aim at creating groups of users that are homogeneous with respect to case-mix. In this way, the group-specific outcomes of a certain agency provide a reliable set of measures of its performance. These can then be suitably aggregated in order to obtain a synthetic performance index. A well-fitting example in the public health sector is represented by the evaluation of Nursing Home (NH) services, where a classification of residents hosted by the NHs based on Resources Utilization Groups (RUGs) has been used for a number of years [2]. Broadly speaking, patients in the same RUG require the same overall care load. Thus, RUGs can be considered as a good proxy of the clinical complexity of each NH resident, thereby representing a useful tool to address case-mix.

In this paper, we focus on a framework where the overall performance of supplying agencies (or a peculiar aspect thereof) can be captured by a binary outcome with 0 denoting failure and 1 denoting success. Sticking to the aforementioned case of NH services, such a variable might indicate whether NH residents died or survived after one year from baseline, though many other examples in diverse fields could be given. Within such a setting, we introduce a general class of case-mix adjusted performance indices encompassing many routinely-used indices as special cases. All indices in this class are built by aggregating group-specific contrasts between the success probability of an agency and the marginal success probability for the whole set of agencies. In this way, the index of an agency can be interpreted in relative terms, that is, as a departure from the average performance. In other words, a comparative performance evaluation is pursued where the average performance is taken as benchmark.

2 A class of adjusted performance indices

Let $h \in \mathscr{H} = \{1,\ldots,H\}$ and $j \in \mathscr{J} = \{1,\ldots,J\}$ be two subscripts indexing supplying agencies and case-mix adjusting groups, respectively. Thus, p_{hj} denotes the success probability for one of the N_{hj} users of the j-th group handled by the h-th agency, whereas $p_{\cdot j}$ denotes the group-specific success probability after marginal-

ization across all agencies. We also introduce the marginal counts $N_{h\cdot} = \sum_{j=1}^{J} N_{hj}$ and $N_{\cdot j} = \sum_{h=1}^{H} N_{hj}$. Under this notation, the generalized Adjusted Performance Index (API) can be written as

$$\text{API}_h^{(G)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj} G(p_{hj}, \boldsymbol{p}), \tag{1}$$

where $\mathscr{J}_h = \{j \in \mathscr{J} : N_{hj} > 0\}$ is the subset of groups for which the agency h handles at least one user, $\boldsymbol{p} = (p_{\cdot 1}, \ldots, p_{\cdot J})$ is the vector of the marginal probabilities for all groups (which of course includes $p_{\cdot j}$) and $W_{hj} = N_{hj}/N_{h\cdot}$ is the proportion of users of group j for agency h.

The $G(p_{hj}, \boldsymbol{p})$ function in (1) defines the nature of the group-specific contrast. In order to obtain a coherent performance measure, such a function is required to have the following properties: i) if $p_{hj} = p_{\cdot j}$ then $G(p_{hj}, \boldsymbol{p}) = 0$; ii) given $\boldsymbol{p}$, $G(p_{hj}, \boldsymbol{p})$ is monotonically non-decreasing with respect to its other argument p_{hj}. These conditions ensure that the resulting index takes values around 1, which occurs when a given agency has a performance equal to the average for every group (that is, $p_{hj} = p_{\cdot j}$ for $j \in \mathscr{J}_h$). Clearly, higher (lower) values denote a better (worse) performance with respect to the average. Moreover, given two agencies h and h' such that $\mathscr{J}_h = \mathscr{J}_{h'}$ and $p_{hj} \geq p_{h'j}$ for every $j \in \mathscr{J}_h$, we have that $\text{API}_h^{(G)} \geq \text{API}_{h'}^{(G)}$ for the monotonicity property of the G function.

The main advantage of the general formulation provided in (1) is a rather high degree of flexibility. Indeed, the G function can be chosen in a number of ways, according to the desired nature of the resulting index. For example, the two functions $G(p_{hj}, \boldsymbol{p}) = p_{hj} - p_{\cdot j}$ and $G(p_{hj}, \boldsymbol{p}) = (p_{hj} - p_{\cdot j})/\sum_{j \in \mathscr{J}_h} W_{hj} p_{\cdot j}$ respectively return

$$\text{API}_h^{(G)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj} p_{hj} - \sum_{j \in \mathscr{J}_h} W_{hj} p_{\cdot j}, \tag{2}$$

and

$$\text{API}_h^{(G)} = \frac{\sum_{j \in \mathscr{J}_h} W_{hj} p_{hj}}{\sum_{j \in \mathscr{J}_h} W_{hj} p_{\cdot j}}. \tag{3}$$

Both (2) and (3) are classical risk-adjusted measures contrasting, for the h-th agency, the marginal success probability ($\sum_{j \in \mathscr{J}_h} W_{hj} p_{hj}$) with the theoretical probability that would result if such an agency had average group-specific probability levels and maintained its case mix distribution ($\sum_{j \in \mathscr{J}_h} W_{hj} p_{\cdot j}$). However, the former operates on the difference scale, while the latter operates on the ratio scale. This example also clarifies why in the generalized formulation (1) every single G function in the summation depends on the whole $\boldsymbol{p}$ vector rather than on the scalar value $p_{\cdot j}$ only.

Another relevant aspect is that different specifications of the G function emphasize different kinds of contrasts. For example, because of some subject-matter reasons one might want a discrepancy between p_{hj} and $p_{\cdot j}$ occurring at higher (lower) values of $p_{\cdot j}$ to outweigh the same discrepancy occurring at lower (higher) values of $p_{\cdot j}$. Also, in certain evaluation contexts it could be desirable that a performance improvement contributes to the index more than a performance worsening of the

same magnitude (or vice-versa). Notice that according to the paradigm the evaluator has in mind these discrepancies might be quantified either as deviations $p_{hj} - p_{\cdot j}$, variations $p_{hj}/p_{\cdot j}$ or relative increments $(p_{hj} - p_{\cdot j})/p_{\cdot j}$. The adoption of a particular paradigm might lead to controversial interpretations of indices tailored to other paradigms; see for example the discussion in [1] concerning indices based on mortality rates. The proposed approach is able to deal with all these instances simply via a sensible specification of the G function.

3 Estimation

Once the G function has been chosen, one has to deal with the problem of estimating the resulting API_h index based on sample data gathered from the agencies. To this end, for each user of group j handled by the h-th agency, we assume we observe a realization from a binary random variable with success probability p_{hj}, and that variables referring to different users are uncorrelated. In this way, for every case-mix adjusting group the observed marginal and agency-specific success rates can be viewed as estimates from unbiased estimators of the corresponding success probabilities. Therefore, the natural estimator of the $\text{API}_h^{(G)}$ index in (1) is

$$\widehat{\text{API}}_h^{(G)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj} G(\hat{p}_{hj}, \hat{\boldsymbol{p}}), \tag{4}$$

where $\hat{p}_{hj}$ e $\hat{\boldsymbol{p}}$ are, in analogy with (1), the sample success rate for the (h, j) pair and the vector of sample marginal success rates.

It is important to highlight that for some groups the finite-sample estimates of the probabilities appearing in the G function might take values on the boundary of the 0-1 interval. In these cases, the index estimate might not be defined depending on the nature of the contrasts and of the transformations involved in the G function (e.g., ratios, logarithms, probability complements). For example, the function $G(\hat{p}_{hj}, \hat{\boldsymbol{p}}) = \hat{p}_{hj}/\hat{p}_{\cdot j} - 1$ leads to the mean of ratios

$$\widehat{\text{API}}_h^{(G)} = \sum_{j \in \mathscr{J}_h} W_{hj} \frac{\hat{p}_{hj}}{\hat{p}_{\cdot j}}, \tag{5}$$

which is not defined when the estimator $\hat{p}_{\cdot j}$ assumes null values for some $j \in \mathscr{J}_h$. When, like in the example above, indefiniteness occurs because of $\hat{p}_{\cdot j}$ taking extreme values for some groups, the problem can be circumvented by setting $G(\hat{p}_{hj}, \hat{\boldsymbol{p}}) \equiv 0$. Indeed, in these cases all agencies have an average performance, be it the best or the worst, and property i) mentioned in Section 2 can be invoked. In what follows, we assume that the probability of an indefiniteness is negligible.

The analytic expressions of the moments of the generalized estimator in (4) can be obtained by linearization, a method offering a good approximation of the true moments when the sample size is large enough. Assuming that $G(\hat{p}_{hj}, \hat{\boldsymbol{p}})$ is differ-

entiable in its domain, a first-order Taylor expansion performed in the point $(p_{hj}, \boldsymbol{p})$ returns

$$\widehat{\text{API}}_h^{(G)} \approx \widetilde{\text{API}}_h^{(G)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj} G(p_{hj}, \boldsymbol{p}) + \sum_{j \in \mathscr{J}_h} W_{hj} A_{hj} (\hat{p}_{hj} - p_{hj}) + \sum_{j \in \mathscr{J}_h} W_{hj} \sum_{\ell \in \mathscr{J}} B_{hj.\ell} (\hat{p}_{.\ell} - p_{.\ell}),$$

where

$$A_{hj} = \left. \frac{\partial G(\hat{p}_{hj}, \hat{\boldsymbol{p}})}{\partial \hat{p}_{hj}} \right|_{\hat{p}_{hj} = p_{hj}, \hat{\boldsymbol{p}} = \boldsymbol{p}} \qquad B_{hj.\ell} = \left. \frac{\partial G(\hat{p}_{hj}, \hat{\boldsymbol{p}})}{\partial \hat{p}_{.\ell}} \right|_{\hat{p}_{hj} = p_{hj}, \hat{\boldsymbol{p}} = \boldsymbol{p}}.$$

It is straightforward to observe that $\widetilde{\text{API}}_h^{(G)}$ is an unbiased estimator of $\text{API}_h^{(G)}$, so that $\widehat{\text{API}}_h^{(G)}$ is asymptotically unbiased for $\text{API}_h^{(G)}$ as $N_{h.} \to \infty$ and $N_{.j} \to \infty$ for $j \in \mathscr{J}$. The variance of $\widetilde{\text{API}}_h^{(G)}$ is given by

$$V\left(\widetilde{\text{API}}_h^{(G)}\right) = \sum_{j \in \mathscr{J}_h} W_{hj}^2 A_{hj}^2 V(\hat{p}_{hj}) + \sum_{j \in \mathscr{J}_h} W_{hj}^2 \sum_{\ell \in \mathscr{J}} B_{hj.\ell}^2 V(\hat{p}_{.\ell}) + 2 \sum_{j \in \mathscr{J}_h} W_{hj}^2 A_{hj} B_{hj.j} \text{Cov}(\hat{p}_{hj}, \hat{p}_{.j}). \tag{6}$$

This result is due to the fact that the random variables $\hat{p}_{hj}$ and $\hat{p}_{h'j'}$ are uncorrelated whenever $h \neq h'$ or $j \neq j'$, which immediately follows from the absence of correlation postulated for the binary variables referring to users. Hence, for every group the only covariance to be accounted for is the one between $\hat{p}_{hj}$ and $\hat{p}_{.j}$. Since $\hat{p}_{.j} = \sum_{h=1}^{H} Q_{hj} \hat{p}_{hj}$, where $Q_{hj} = N_{hj}/N_{.j}$ is the share of users of group j handled by the h-th agency, we have that such a covariance is $\text{Cov}(\hat{p}_{hj}, \hat{p}_{.j}) = Q_{hj} V(\hat{p}_{hj})$. Substituting in (6) returns a more compact (approximate) expression for the target variance which, again, holds true asymptotically and is given by

$$V\left(\widehat{\text{API}}_h^{(G)}\right) \approx \sum_{j \in \mathscr{J}_h} W_{hj}^2 A_{hj} (A_{hj} + 2 Q_{hj} B_{hj.j}) V(\hat{p}_{hj}) + \sum_{j \in \mathscr{J}_h} W_{hj}^2 \sum_{\ell \in \mathscr{J}} B_{hj.\ell}^2 V(\hat{p}_{.\ell}), \tag{7}$$

where

$$V(\hat{p}_{hj}) = \frac{p_{hj}(1 - p_{hj})}{N_{hj}} \qquad V(\hat{p}_{.\ell}) = \sum_{h=1}^{H} Q_{h\ell}^2 \frac{p_{h\ell}(1 - p_{h\ell})}{N_{h\ell}}. \tag{8}$$

For instance, the approximate variance of the estimator in (5) is

$$V\left(\widehat{\text{API}}_h^{(G)}\right) \approx \sum_{j \in \mathscr{J}_h} W_{hj}^2 \left\{ \frac{p_{.j} - 2 Q_{hj} p_{hj}}{p_{.j}^3} V(\hat{p}_{hj}) + \frac{p_{hj}^2}{p_{.j}^4} V(\hat{p}_{.j}) \right\},$$

with $V(\hat{p}_{hj})$ and $V(\hat{p}_{\cdot j})$ as in (8). A consistent estimator of the right-hand side of (7) is obtained by plugging-in $\hat{p}_{hj}$ and $\hat{p}_{\cdot\ell}$ in place of the true probabilities p_{hj} and $p_{\cdot\ell}$ appearing in the A_{hj} and $B_{hj\cdot\ell}$ terms as well as in the two variances in (8), for which the usual sample size corrections (that is, replacing N_{hj} and $N_{h\ell}$ with $N_{hj}-1$ and $N_{h\ell}-1$, respectively) can also be implemented in order to achieve unbiased estimation.

4 Conclusions

Like other tools dealing with relative performance measuring, the proposed family of indices is likely to be mainly used for comparison and/or ranking purposes. Clearly, the inferential framework sketched in Section 3 allows the evaluator to corroborate point-estimation results with their statistical significance. In this regard, the general expression of the covariance between the indices of two different agencies can be derived to draw inferential conclusions on pairwise differences.

Further research is ongoing on the class of case-mix adjusted probability-based indices for performance evaluation proposed in this work. For instance, the finite-sample properties of the estimators associated to the most popular indices (including, but not limited to, those presented in the previous sections) should be studied via simulation. Indeed, small sample sizes are likely to occur in many applied settings, posing a number of challenges which range from the tenability of first-order Taylor approximations to the impossibility of estimating the variance of a sample proportion when $N_{hj}=1$. For these issues, suitable remedies are being studied.

Acknowledgements We are thankful to Cassa di Risparmio di Perugia for financial support and to Regione Umbria (Italy) for sharing the data that inspired this work.

References

1. Castro, R.A.S., Oliveira, P.N., Silva Portela, C., Camanho, A.S., Melo, J.Q.: Benchmarking clinical practice in surgery: looking beyond traditional mortality rates. Health Care Manag. Sc. **18**, 431–443 (2015)
2. Fries, B.E., Schneider, D.P., Foley, W.J., Gavazzi, M., Burke, R., Cornelius, E.: Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). Med. Care **32**, 668–685 (1994)
3. Kruk, M.E., Freedman, L.P.: Assessing health system performance in developing countries: A review of the literature. Health Policy **85**, 263–276 (2008)

An ultrametric model to build a Composite Indicators system

Un modello ultrametrico per costruire un sistema di indicatori compositi

Carlo Cavicchia, Pasquale Sarnacchiaro, Maurizio Vichi and Giorgia Zaccaria

Abstract In the last years, the use of composite indicators has consistently increased, and the necessity to build model-based composite indicators with a strong methodological statistical approach becomes more and more important for reasons of trustworthiness. In this paper, we propose to build a composite indicators system able to measure different levels of relations among (group of) variables according to an ultrametric form which detects a hierarchical structure upon (group of) variables. Each dimension is measured as a specific composite indicator which reflects a subset of variables. In order to show its potential and applicability, the methodology is employed to analyze a dataset which contains variables about separated waste collection in Italy taking into consideration both its performance and its costs.

Abstract *Negli ultimi anni l'utilizzo di indicatori compositi è costantemente cresciuto, e la necessità di costruire degli indicatori compositi model-based con un forte approccio statistico è sempre più importante per motivi di fiducia. In questo articolo proponiamo di costruire un sistema di indicatori compositi che possa misurare diversi livelli di relazioni tra (gruppi di) variabili seguendo una forma ultrametrica che individui una gerarchia sulle (gruppi di) variabili. Al fine di mostrare il suo potenziale e la sua applicabilità, la metodologia è applicata per analizzare*

Carlo Cavicchia
Erasmus University Rotterdam, Rotterdam, The Netherlands
e-mail: cavicchia@ese.eur.nl

Pasquale Sarnacchiaro
University of Naples Federico II, Naples, Italy
e-mail: sarnacch@unina.it

Maurizio Vichi
University of Rome La Sapienza, Rome, Italy
e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome Unitelma Sapienza, Rome, Italy
e-mail: giorgia.zaccaria@unitelmasapienza.it

un dataset che contiene variabili riguardo la raccolta differenziata in Italia considerando sia le sue prestazioni che i suoi costi.

Key words: Latent variable model, Hierarchical model, Model-based, Latent concept, Statistical estimation

1 Introduction

Composite Indicators (CIs) are non-observable latent variables which consist of the aggregation of observed variables into a single non-observable index according to an underlying model for the multidimensional concepts [7, 8]. A CI is therefore a mathematical (weighted) combination of variables that generally is subject to several choices by the researcher [2]. CIs are able to summarize a big amount of information and for this specific feature they are very useful to measure multidimensional phenomena by potentially highlighting different levels of synthesis. However, the methods for CIs' construction are often criticized since they are not considered statistically rigorous or based on theories with solid foundations [6], thus, they might lead to misleading results and interpretations. This accounts for building CIs via a model-based approach.

In this paper, we propose to build a CIs system able to measure different levels of relations among (group of) observed variables according to an ultrametric structure [3, 4]. This structure allows describing multidimensional phenomena which are characterized by nested latent concepts having different levels of abstraction, from the most specific to the most general. In detail, internal consistent concepts are built and eventually aggregated from the most concordant ones to the most discordant. The proposal therefore detects a hierarchical structure upon variables.

In order to show its potential and applicability, the methodology is employed to analyze a dataset which contains variables about separated waste collection in Italy taking into consideration both its performance and its costs. This topic results crucial nowadays since many States are still land-filling huge amounts of municipal waste – the worst waste management option – despite the existence of better alternatives, and notwithstanding structural funds being available to finance better options. It is thus worth investigating how to measure the goodness and affordability of a waste management service via a system of CIs which assess each latent concept included in its definition. The number of information and statistics about waste management is larger and larger. For instance, Cavicchia, Sarnacchiaro and Vichi [1] detected which dimensions have an impact on the waste management in EU building a general composite indicator based on three specific composite indicators: recycling and circular economy performance, generation of recyclable waste, and private investments and innovation.

2 Methodology

CIs are able to reduce complex phenomena to a unique measure which results easier to interpret and might be used during the policy-making process. Notwithstanding their usefulness, it might be important to use both a set of specific indicators and a unique aggregated index. This means that a complex reality must be represented at different levels of abstraction and synthesis which might help to understand better the specific characterizations of the phenomenon that is being studied. The aggregated CI is the result of an entire hierarchy which starts from Q internal consistent latent concepts, in turn the hierarchy is provided by the ultrametric structure that reconstructs the main relationships among the variables. In other words, the hierarchy is composed of nested dimensions characterized by distinct levels of abstraction.

The different levels of the hierarchy are reconstructed thought four matrices: 1) a $p \times Q$ membership matrix $\mathbf{V}$, which represents the membership of each variable to a group where p is the number of the observed variables; 2) a diagonal matrix $\mathbf{S}^V$ of order Q, whose main diagonal represents the variance of each group; 3) a diagonal matrix $\mathbf{S}^W$ of order Q, whose main diagonal represents the covariance within each group; 4) a ultrametric matrix $\mathbf{S}^B$ of order Q, whose diagonal entries are set to zero and off-diagonal ones represent the hierarchical relationships among pairs of concepts. Whereas the CIs are built as the score vectors which best reconstruct the data matrix. The model-based approach which characterizes this hierarchy and this system of CIs guarantees to optimize an objective function in the least-squares framework.

The proposal extends the work by Cavicchia, Vichi and Zaccaria [5] by also reconstructing the covariance structure of the observed variables via an extended ultrametric covariance matrix [4]. However, the proposed method preserves the feature to obtain a reduced number of latent concepts which are quantified by maximizing the explained variance. These two goals are reach by minimizing a common objective function.

3 Application

The data used in this application about the separated waste collection are from different sources: Eurostat, Joint Research Centre (JRC) and Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA). It is worth observing that the observed variables used in this paper - namely, Cost of separated waste collection and transport, Cost of separated waste treatment and recycle, Organic waste collection, Paper waste collection, Glass waste collection, Metal waste collection, Plastic waste collection and Percentage of separated waste over the total waste - are regularly updated and free. Two variables represent the costs of the separated waste while the other six variables express the performance of it. In detail, we include in our analysis only the largest 40 Italian municipalities for comparability reason and we use the size population (i.e., number of inhabitants) and the size of waste produced (i.e., weight in

kilograms) to normalize the variables yet Percentage of separated waste over the total waste. This choice allows us to conduct two different analyses: one regarding the efficiency of the separated waste collection and one regarding the efficacy. A few steps of pre-processing are taken into consideration: the few missing data are Missing Completely at Random (MCAR) and therefore imputed by the K-nearest neighbors method by setting $K = 4$ and by using the Euclidean distance; and the variables are then standardized.

The motivation of this study lies on the assumption that it is crucial to combine the information from the costs and the performance to provide a support for Italian municipalities' actions and policies. The information from these two aspects, if measured separately, might be either misleading or limited. The research aims at searching other important latent concepts which might be present withing the main two, namely, costs and performance.

4 Conclusion

This paper provides a model-based approach to build a CIs system able to pinpoint a hierarchy and the quantification of the latent concepts which compose it. Furthermore, this study presents a useful tool to measure the separated waste collection in Italy together with its main aspects, by identifying the most important relationships among variables. The goal is to provide both a methodological contribution to the construction of CIs literature and a support for Italian municipalities' actions and policies.

References

1. Cavicchia, C., Sarnacchiaro, P. and Vichi, M.: A composite indicator for the waste management in the EU via hierarchical disjoint non-negative factor analysis. Socio-Economic Planning Sciences **73**, 100832 (2021)
2. Cavicchia, C. and Vichi, M.: Statistical Model-based Composite Indicators for tracking coherent policy conclusions. Social Indicators Research **156(2)**, 449-479 (2021)
3. Cavicchia, C., Vichi, M. and Zaccaria, G.: The ultrametric correlation matrix for modelling hierarchical latent concepts. Advances in Data Analysis and Classification **14(4)**, 837-853 (2020)
4. Cavicchia, C., Vichi, M. and Zaccaria, G.: Gaussian Mixture Model with an extended ultrametric covariance structure. Submitted. (2021)
5. Cavicchia, C., Vichi, M. and Zaccaria, G.: Hierarchical Disjoint Principal Component Analysis. Submitted. (2021)
6. Mazziotta, M. and Pareto, A.: Methods for constructing composite indices: One for all or all for one? Rivista Italiana di Economia Demografia e Statistica **67(2)**, 67-80 (2013)
7. Nardo, M., Saisana, M., Saltelli, A. and Tarantola, S.: Tools for Composite Indicators Building. European Commission (Join Research Centre, Ispra, Italy). Report EUR 21682. (2015)
8. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E.: Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing. OECD Statistics Working Papers 2005/3 (2005)

Session of free contributes SCL6 – *Statistical Learning*
Chair: Massimo Aria

Feature definition for NBA result prediction through Deep Learning

Definizione delle features per la predizione dei risultati nella NBA tramite il Deep Learning

Manlio Migliorati, Eugenio Brentari

Abstract This contribution is focused on features' definition for the outcome prediction of matches of NBA basketball championship. It is shown how models based on one a single feature (Elo rating or the relative victory frequency) can have a quality of fit better than models using box-score predictors (e.g. the Four Factors). Features have been ex ante calculated for a dataset containing data of 16 NBA regular seasons, paying particular attention to home court factor. Models have been produced via Deep Learning, using cross validation.

Abstract *Questo contributo è focalizzato sulla costruzione di predittori per la predizione dei risultati degli incontri del campionato di basket NBA. In particolare si mostra come modelli basati su un unico predittore (Elo rating o la frequenza relative delle vittorie) possono avere una qualità di fit superiore a quella dei modelli basati sui box-scores (ad esempio i Four Factors). I predittori sono stati calcolati ex-ante su un dataset che comprende i dati di 16 regular seasons del campionato NBA, facendo particolare attenzione al fattore campo. I modelli sono stati prodotti tramite Deep Learning, applicando la cross-validation.*

Key words: basketball outcome prediction, features definition, court factor

1 Introduction

This contribution is focused on features selection for the problem of predicting the winner in NBA matches. It is shown how, for outcome prediction classification problem, a careful definition of single features used in model definition, can produce predictions with a quality better than quality of models built on the top of box-score statistics.

University of Brescia, Department of Economics and Management, Contrada S. Chiara 50
e-mail: manlio.migliorati@unibs.it
e-mail: eugenio.brentari@unibs.it

To this purpose, two features directly quantifying strength of teams involved in a match have been selected:

1. The Elo (from the name of its creator) rating system [2], originally defined for rating chess players and today widely used in several domains.
2. The difference of the relative frequency of victories for the two teams.

and used as covariates to build models to be compared, in terms of quality of fit, to models built using Oliver's Four Factors [5, 4] (few indexes synthesizing several *box-score* statistics) as regressors.
The models built in this work have been developed in a particular Deep Learning echosystem in R based on `Keras` package [1].

2 Features' definition

2.0.1 The Elo rating

The Elo rating system [2] has been originally defined for calculating the strength of players in zero-sum games (i.e. games where a player gains exactly what its opponent loses) as chess, the sport for which this system was created by Arpad Elo. More formally: if before a match Player1 has a rating $R1$ and Player 2 has a rating $R2$, and if S is the result of the match (where 1 means Player1 victory and 0 Player2 victory), after the match, the ratings of the 2 players will be updated as follows:

$$R1^{'} = R1 + K * (S - P(p1w)) \tag{1}$$

$$R2^{'} = R2 + K * (S - P(p2w)) \tag{2}$$

where K is a parameter addressing how strongly a result will affect ratings' update and $P(p1w)$ and $P(p2w)$ are the probabilities of victory (modelled as logistic curves) attributed to the two players before the match. The difference in Elo ratings between the two teams fighting in a match will be the first feature used in the present study: for the initial ratings we will follow [6] and 1300 will be used.

2.0.2 The difference in relative victory frequencies

A second feature directly quantifying the strength of opposing teams is the difference of their relative victory frequencies, named *diff* in the following. It can be formally defined as follows:

$$diff = \frac{won\_matches_{ht}}{played\_matches_{ht}} - \frac{won\_matches_{at}}{played\_matches_{at}} \tag{3}$$

Where the subscript *ht* means home team, and the subscript *at* mean away team. *Diff* statistics ranges from -1 to 1, where value 1 means that the home team is absolutely

the strongest between the two teams. So, *diff* is a clear and concise way for showing the difference in class between the two teams, providing an analytical definition for a classic rule of thumb often used in naive fan predictions (the favorite is the team that won more in the (recent) past).

2.0.3 Four Factors

The Four Factors [5, 4] is a set of indexes built on top of classic *box-score* statistics. Four Factors are considered fundamental for winning a match, and summarize the attitude of a team with respect to shooting, turnovers, rebounds and free throws.

3 The dataset

The dataset includes data about 16 NBA regular seasons (from 2004-2005 to 2019-2020), counting more than 18.000 observations (one for each match). Elo, $diff$ and Four Factors have been calculated ex ante, i.e. considering only information from prior matches, to make them suitable for outcome predictions, taking into account:

- the periodicity, considering both the historical (considering all prior games) and the dynamic perspective (averaging on a subset of prior matches). Moreover, the mechanism of regression to mean [3] has been implemented for historical features, seeming particularly suitable for NBA [6], where at the end of each season there is an attempt to rebalance teams strength.
- the court where matches have been played: besides features usually calculated considering all matches, two new statistics based considering only either home or away data (called *the court issue* in the following) will be calculated, too.

4 Methods and Models: Deep Learning

4.1 Building Deep Learning models

All the models described in this work share the same sequential structure:

- one first input layer, with a number of input units corresponding to the number of features to be considered in building the model (1 for Elo and $diff$, 8 for Four Factors (4 for each team)
- one final output layer, with 1 output unit corresponding to the two possible results of a NBA match (basketball outcome prediction is a typical classification problem)

- a stack of several intermediate hidden sequential layers, connecting the input and output layers. Each hidden layer contains several elaboration units, to work on data received from the prior layer before sending them to the following layer.

The nets, calibrated to produce models with a good prediction quality, are built considering the two hyperparameters (i.e. the number of layers and the number of units for each layer) small in size, a natural consequence of the small number of features.

5 Results

The results reported in this section have been obtained using a v-fold cross-validation with v=4.

5.1 Using Elo features

Execution results for models based on Elo variants are reported in Table 1. The quality of predictions for models built using historical Elo without considering the court issue is the best one, with an AUC equal to 0.7117 and an accuracy equal to 0,6721 (using a threshold equal to 0.5047). These values have been obtained using a regression to mean percentage *P%* equal to 20.
Between the models built using dynamic Elo, the model not considering the court issue, obtained with a depth equal to two, is the best one: its AUC is equal to 0.7117 and its accuracy equal to 0.6736 (threshold equal to 0.5049), the best among the models we built in this work. Also predictions' quality for the model built using dynamic Elo considering the court issue, obtained with a depth equal to three, is good, with an AUC equal to 0.7103 and an accuracy equal to 0.6705 (threshold equal to 0.5148).

Table 1 Best quality of predictions for models based on Elo. For each variant, the best AUC measure, the corresponding threshold and the accuracy measure are reported, together with parameters' values used in Elo calculation

periodicity	court issue	AUC	threshold	accuracy	regression to mean P%
historical	not considered	0.7117	0.5047	0.6721	20
historical	considered	0.7001	0.5058	0.6650	60
periodicity	court issue	AUC	threshold	accuracy	depth
dynamic	not considered	0.7117	0.5049	0.6736	2
dynamic	considered	0.7103	0.5148	0.6705	3

5.2 Using $diff$ features

Results are reported in Table 2. The quality of predictions of the model built using $diff$ without considering the court issue is the best one, with an AUC equal to 0.6925 and an accuracy equal to 0.6626 (using a threshold equal to 0.5236). For the model built using dynamic $diff$, the quality of predictions not considering the court issue is the best one, with an AUC equal to 0.7020 and an accuracy equal to 0,663 (threshold equal to 0.5255).

Table 2 Best quality of predictions for models based on $diff$. For each variant, the best AUC measure, the corresponding threshold and the accuracy measure are reported, together with parameters' values used for calculation

periodicity	court issue	AUC	threshold	accuracy	regression to mean P%
historical	not considered	0.6925	0.5236	0.6626	90
historical	considered	0.6775	0.4788	0.6572	78
periodicity	court issue	AUC	threshold	accuracy	depth
dynamic	not considered	0.7020	0.5255	0.663	50
dynamicl	0.6944	0.5057	0.6586	27	

5.3 Using Four Factors

Table 3 reports some results: the model built on historical Four Factors without considering the court issue is the best one, with an AUC equal to 0.6655 and an accuracy equal to 0.6400 (threshold equal to 0.5334). Between dynamic features, the two models are equivalent in terms of quality of fit, slightly less than quality of historical model.

Table 3 Best quality of predictions for models based on Four Factors. For each variant, the best AUC, the corresponding threshold and the accuracy measure are reported, together with the parameter's value used for calculation

periodicity	court issue	AUC	threshold	accuracy	regression to mean P%
historical	not considered	0.6655	0.5334	0.6400	78
historical	considered	0.6527	0.4968	0.6347	74
court issue	AUC	threshold	accuracy	depth %	
dynamic	not considered	0.6495	0.4934	0.6371	42
dynamic	considered	0.6492	0.5091	0.6372	36

6 Conclusions

In this contribution we showed how appropriately defined statistics can profitably be used as single features in fitting models for outcome predictions on a basketball dataset including 16 NBA regular seasons from 2004-2005 to 2019-2020.
The models quality is better than quality of models fitted using Four Factors, a synthesis of *box-score* statistics.
The best prediction quality for a model considering the whole period has been produced using a single dynamic Elo feature (not considering the court issue), with an averaging depth equal to two (i.e. only Elo rating of prior two matches are considered in feature calculation). For this model, the AUC is equal to 0.7117 and the accuracy (using a threshold equal to 0.5049) is equal to 0.6736 (same AUC of the model built using historical Elo, but higher accuracy).
Results suggest that the court issue approach to features definition produces predictions comparable in the quality to models based on usual single feature, offering more interpretation details. Moreover, we verified how regression to mean can play a relevant role in prediction quality.
In general, quality of models built using $diff$ based features is close to quality of models built using Elo, and this is an expected result if we take into account how both these features express a direct measure of the strength of a team. Instead, the quality of models based on Four Factors is remarkably the lowest among the three approaches, suggesting how the approaches based on *box-score* statistics are close to their limit in outcome prediction quality.

Acknowledgements We would like to thank Prof. Marica Manisera for the help she gave us during this work.

References

1 Allaire, J., Chollet, F.: keras: R Interface to Keras (2021). URL https://CRAN.R-project.org/package=keras
2 Elo, A.E.: The Rating of Chess players, Past and Present. Ishi Press International (1978)
3 Galton, F.: Natural Inheritance. MacMillan (1889)
4 Kubatko, J., Oliver, D., Pelton, K., Rosenbaum, D.: A starting point for analysing basketball statistics. Journal of Quantitative Analysis in Sports **3**(3), 1–22 (2007)
5 Oliver, D.: Basketball on Paper: Rules and Tools for Performance Analysis. Potomac Books inc (2004)
6 Silver, N.: How We Calculate NBA Elo Ratings (2015). URL https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

An application of contrast trees for mortality models diagnostic and boosting

Un'applicazione degli alberi di contrasto per la diagnostica e il boosting di modelli di mortalità

Susanna Levantesi, Matteo Lizzi and Andrea Nigri

Abstract Different mortality models are hardly comparable with the observed data in a unified framework. In this paper, we use contrast trees as a general method that allows us to assess and compare the quality of different morality models, fitted on the Italian mortality rates. The results are discussed using both graphical and numerical tools, with particular attention to the low-performing regions.

Abstract *Diversi modelli di mortalità sono difficilmente confrontabili con i dati osservati in un quadro unificato. In questo articolo, utilizziamo alberi di contrasto come metodo generale che ci consente di valutare e confrontare la qualità di diversi modelli di moralità, stimati su tassi di mortalità italiani. I risultati vengono discussi utilizzando strumenti sia grafici che numerici, con particolare attenzione alle regioni con alti livelli di errore.*

Key words: mortality modeling, machine learning, contrast trees

1 Introduction

The longer life experienced in the modern era represents an enormous achievement but at the same time it poses major challenges. Estimating longevity is not straightforward: it is not clear when a method will perform best. As proposed by Booth and Tickle (2008) the accuracy of population estimates should be regularly tested. Until the 1980s, mathematical mortality models were relatively simple and involved a fair degree of subjective judgment. The availability

Susanna Levantesi · Matteo Lizzi
Department of Statistics, Sapienza University of Rome, Viale Regina Elena 295-G, 00161 Rome, e-mail: susanna.levantesi@uniroma1.it,matteo.lizzi@uniroma1.it

Andrea Nigri
Department of Social and Political Sciences, Bocconi University, Milan, Italy
e-mail: andrea.nigri@unibocconi.it

of reliable data, in lockstep with the improvement of statistical-mathematical methods, has allowed the creation of ever-finer mortality forecasting models. According to Booth and Tickle (2008), literature suggests three approaches to demographic modeling: *explanation* makes use of structural models of mortality from certain causes of death, *expectation*, based on expert opinion, involving varying degrees of formality. Finally, *extrapolation* uses the regularity found in age patterns and trends over time (Lee and Carter, 1992). Using a new method proposed by (Friedman, 2020), namely Contrast Trees, this paper offers different a perspective for evaluating the accuracy of the mortality estimates provided by machine learning algorithms.

2 Methodology

The goal of the Contrast Trees (CTs) method is to uncover regions in the predictor variables space with very high values of the error rate quantified by a discrepancy measure (Friedman, 2020). CTs are easy to be interpreted and can be used as diagnostic tools to detect the inaccuracies of models engendered by any learning method.

Suppose to have a set of predictor variables $x = (x_1, x_2, ..., x_p)$ and two outcome variables y and z for each x. We aim to find those values of x for which the respective distributions of $y|x$ and $z|x$, or some statistics such as mean or quantiles, are most different. In summary, CTs provide a lack-of-fit measure for the conditional distribution $p_y(y|x)$, or some statistics.

Let consider the M^{th} iteration, where the tree splits the space of the predictor variables into M disjoint regions $\{R_m\}_{m=1}^{M}$, each one containing a subset of the data. Let denote $f_m^{(l)}$ and $f_m^{(r)}$ the fraction of observations in the left and right region with respect to R_m, respectively. The quantities $d_m^{(l)}$, $d_m^{(r)}$ respectively represent the discrepancy measures associated to the fractions $f_m^{(l)}$ and $f_m^{(r)}$. Given a specified subset of the data $\{x_i, y_i, z_i\}_{x_i \in R_m}$, a discrepancy measure between y and z values can be generally defined as:

$$d_m = D\left(\{y_i\}_{x_i \in R_m}, \{z_i\}_{x_i \in R_m}\right) \tag{1}$$

The quality of a split is quantified by the following measure:

$$Q_m(l, r) = \left(f_m^{(l)} \cdot f_m^{(r)}\right) \cdot max\left(d_m^{(l)}, d_m^{(r)}\right)^{\beta} \tag{2}$$

The choice of the discrepancy measure depends on the problem to be solved, allowing CTs to be applied to a variety of problems. In numerical applications, for sake of simplicity the following discrepancy measure $d_m^{[1]}$ has been used:

$$d_m^{[1]} = \frac{1}{N_m} \sum_{x_i \in R_m} |y_i - z_i| \tag{3}$$

where N_m is the number of observations in the region R_m. See Friedman (2020) for further details about the tree split procedure.

2.1 Estimation boosting

To improve the models, Friedman (2020) proposes a contrast-boosting strategy that, dealing with the uncovered errors, can enable the regression models to provide more accurate predictions. Contrast boosting works by gradually modifying a starting value of z to reducing its discrepancy with y over the data. The resulting prediction is then affected by these modifications on the initial value of z. Specifically, we consider the estimation contrast boosting, which takes z as an estimate of a parameter of the full conditional distribution of a target variable given a set of predictor variables, $p_y(y|x)$.

The procedure consists of modifying the z values within a certain region $R_m^{(1)}$ of a CT, so that its discrepancy with y is zero, i.e. to set $d_m = 0$ in Eq. 1. This is an iterative procedure, where the first modification is from z to $z^{(1)} = z + \delta_m^{(1)}$ for $x \in R_m^{(1)}$, the second from $z^{(1)}$ to $z^{(2)} = z + \delta_m^{(2)}$ for $x \in R_m^{(2)}$, and so on. The z values final estimate is then $\tilde{z}(x) = z(x) + \sum_{k=1}^{K} \delta_m^{(k)}$, where K are the maximum number of iterations. In practice, each updated value of z is contrasted with y producing new regions $R_m^{(k)}$ $(1 \leq k \leq K)$ with corresponding updates $\delta_m^{(k)}$.

3 Numerical Application

We consider the Italian mortality data available in the Human Mortality Database (HMD) over the period 1950-2018, referred to male population aged 30-90, and analyzing separately age groups 30-60 and 61-90. We split the data into training set and test set according to the common rule 70%-30%. The dataset partition is obtained by using the dissimilarity-based compound selection proposed in Willett (1999). We apply the methodology to the following models in a regression framework:

- *Lee-Carter (LC) model.* The model is specified as:

$$\log(m_{x,t}) = \alpha_x + \beta_x \kappa_t \tag{4}$$

We assume that deaths are independent Poisson distributed (Brouhns et al., 2002). The LC model can be reformulated in a GNM framework according to the Poisson assumption for death counts.

- *Age-Period-Cohort (APC).* We use the model version reformulated in a GLM framework (Alai and Sherris, 2014):

$$\log(m_{x,t}) = \beta_0 + \beta_{1,x} + \beta_{2,t} + \beta_{3,x+t} \tag{5}$$

- *Gradient boosting machine (GBM).* A tree-based algorithm proposed by Friedman (2001) that uses fixed size decision trees as weak learners.
- *Extreme gradient boosting (XGBM).* An efficient implementation of gradient boosting decision trees, proposed by Chen et al. (2015)and applied to both raw and preprocessed data: the latter is obtained by centering and scaling.

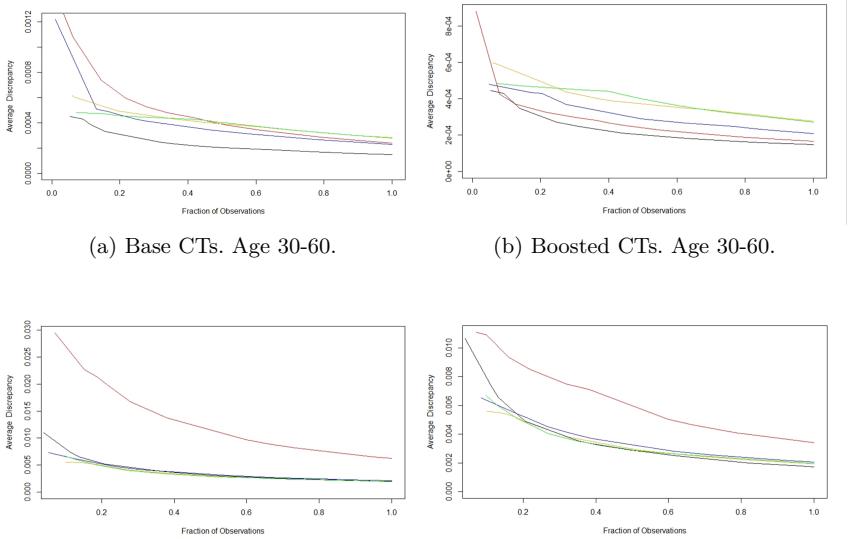
3.1 Results

The analyses have been implemented via the *conTree* R package (Friedman and Narasimhan, 2020)setting thr maximum tree size to 100. T In the following, we will show how CTs can be used to assess different models. We first summarize the results using the lack-of-fit contrast curves, which have point coordinates $[f_m, \bar{d}_m]$. $f_m = \frac{1}{N}\sum_{d_j \geq d_m} N_j$ is the fraction of observations in the Region R_m containing N_m observations, while $\bar{d}_m = \dfrac{\sum_{d_j \geq d_m} d_j N_j}{\sum_{d_j \geq d_m} N_j}$ is the weighted average discrepancy. The lack-of-fit contrast curves obtained by contrasting the observed data to the estimates provided by each models are reported in the left panels of Fig. 1, while those comparing accuracy estimates after applying contrast boosting to the output of the models, are illustrated in the right panels. Table 1 reports the values of the average discrepancy measure for both the base and the boosted models considered in the analysis.

Table 1: Average discrepancy measure

Model	Age 30-60		Age 61-90	
	Base	Boosted	Base	Boosted
APC	1.50E-04	1.47E-04	2.01E-03	1.74E-03
LC	2.29E-04	2.08E-04	2.13E-03	2.07E-03
GBM	2.41E-04	1.67E-04	6.23E-03	3.40E-03
XGBM	2.83E-04	2.78E-04	1.96E-03	1.95E-03
XGBM prep	2.81E-04	2.72E-04	1.97E-03	1.97E-03

As could be expected from the left panel of figure 1, for age group 30-60 the APC model has the lowest average discrepancy. However, lack-of-fit curves



(a) Base CTs. Age 30-60.

(b) Boosted CTs. Age 30-60.

(c) Base CTs. Age 61-90.

(d) Boosted CTs. Age 61-90.

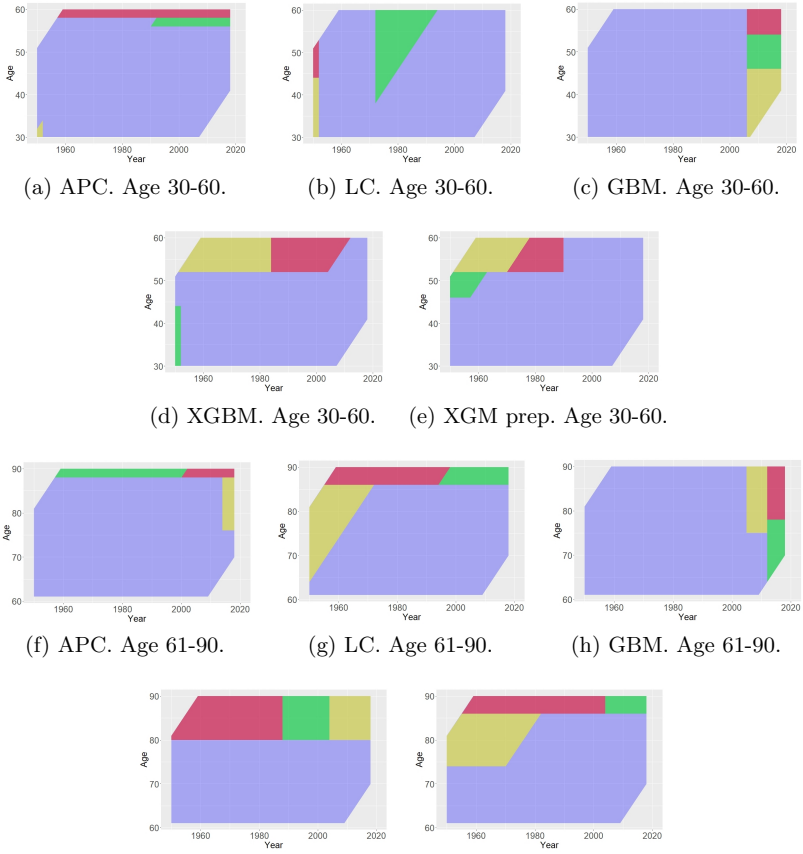
Fig. 1: Lack-of-fit contrast curves for APC (black), LC (blue), GBM (red), XGBM (orange) and XGBM prep (green)

provide more structured information, particularly regarding how and how much model discrepancy varies across the regions identified in input space.

These regions can be identified and interpreted. Moreover, high-discrepancy regions can be used for assessing whether or where a model should be trusted or not. For sake of brevity we show in Fig. 2 the three highest-error regions.

References

1. Alai, D.H., Sherris, M.: Rethinking age-period-cohort mortality trend models. Scand. Act. J. **3**, 208–227 (2014)
2. Booth, H., Tickle, L.: Mortality modelling and forecasting: A review of methods. Ann. Act. Sci. (2008) doi: 10.1017/S1748499500000440.
3. Brouhns, N., Denuit, M., Vermunt, J.: A Poisson log-bilinear approach to the construction of projected life tables, Ins. Math. Econ. **31**, 373–393 (2002)
4. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H.: Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1–4 (2015)



(a) APC. Age 30-60. (b) LC. Age 30-60. (c) GBM. Age 30-60.

(d) XGBM. Age 30-60. (e) XGM prep. Age 30-60.

(f) APC. Age 61-90. (g) LC. Age 61-90. (h) GBM. Age 61-90.

(i) XGBM. Age 61-90. (j) XGM prep. Age 61-90.

Fig. 2: Contrast tree regions. Age 30-60 above, age 61-90 below. The three highest-error regions (colored in red, yellow, and green in decreasing error order).

5. Friedman, J.H.: Greedy function approximation: A Gradient Boosting Machine. Ann. Stat. **29**, 1189–1232 (2001)
6. Friedman, J.H.: Contrast trees and distribution boosting. Proc. Natl. Acad. Sci. (2020) doi: 10.1073/pnas.1921562117
7. Friedman, J.H., Narasimhan, B.: conTree: Contrast Trees and Distribution Boosting. R package version 0.2-8 (2020)
8. Lee, R.D., Carter, L.R.: Modeling and forecasting US mortality. J. Am. Stat. Ass. **87**, (419) 659–671 (1992)
9. Willett, P.: Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. J. Comp. Biol. **6**, (3-4) 447–457 (1999)

Twenty Years of Random Forest: preliminary results of a systematic literature review

Venti anni di Random Forest: una review sistematica preliminare

Massimo Aria, Agostino Gnasso and Luca D'Aniello

Abstract The Random Forest (RF) model consists of an ensemble classifier that produces many decision trees through the use of a randomly selected subset of samples and training variables. The RF model has assumed importance within the scientific community thanks to its performance. The accuracy of its classifications and prediction has allowed the use of RF in several research domains, which have benefited from it. The present study aims to provide a preliminary review of the whole scientific production characterized by all the publications citing the article "Random Forest" by Breiman, 2001, in the last 20 years (2001-2021).

Abstract *L'approccio Random Forest (RF) consiste in un classificatore di ensemble che produce un grande numero di alberi decisionali, attraverso l'uso di un sottoinsieme di variabili, casualmente selezionato. Il modello RF ha assunto importanza all'interno della comunità scientifica grazie alle sue prestazioni, nonchè granzie all'accuratezza delle sue classificazioni e previsioni. Il presente studio mira a fornire un'antemprima preliminare dell'intera produzione scientifica caratterizzata da tutti i lavori accademici che hanno citato l'articolo "Random Forest" di Breiman, 2001, negli ultimi 20 anni (2001-2021).*

Key words: Random Forest, Bibliometrics, Systematic literature review, Science mapping

Massimo Aria
Department of Economics and Statistics, University of Naples Federico II, Italy
email: massimo.aria@unina.it

Agostino Gnasso
Department of Economics and Statistics, University of Naples Federico II, Italy
email: agostino.gnasso@unina.it

Luca D'Aniello
Department of Social Sciences, University of Naples Federico II, Italy
email: luca.daniello@unina.it

1 Introduction

Machine learning is a data analysis approach that automates the construction of analytical models. It is a branch of Artificial Intelligence based on the idea that systems can learn from data, identify patterns on their own and make decisions with minimal human intervention [10]. The use of Ensemble methods in Machine Learning provides different predictive models with different results for the same inputs. Ensemble Learning approaches increase predictive performance models by combining the outputs of a set of induced hypotheses, also called base learners, into a single predictive model. It has the purpose of decreasing variance, altering bias, and improving predictions. An ensemble learner can match any machine learning algorithm such as the decision tree, neuronal network, or a linear regression model. Classification and Regression Trees (CART) are supervised learning techniques that use a nonparametric approach [5]. The process of building trees is intuitive and simple for the human mind, which implies a simple and useful interpretation, but it is not competitive in terms of accuracy concerning other regression and classification approaches. However, predictive performance can be substantially improved by aggregating many decision trees.

In 2001 Leo Breiman proposed Random Forest (RF) method [4], a non-linear approach aims to achieve greater accuracy by averaging multiple decision trees, each of which is grown according to two random steps: the first step consist of the use of a bootstrap sample to train each tree, while the second is the use, at each internal node, of a random subset of variables to generate splits. RF is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set and aggregates predictions on a new record [3].

The purpose of this work is to present a systematic literature review of the last twenty years – from the publishing of RF papers to date - and identify the main research domain that use RF method through quantitative and longitudinal analysis.

2 Materials and methods

Bibliometrics has the potential to introduce a systematic, transparent, and reproducible review process based on the statistical measurement of science, scientists, or scientific activity [6] [12]. Bibliometric analysis involves quantitative methods for exploring, monitoring, and measuring of published research with a set of tools within one or more specified fields over a given period of time [15].

This work aims to perform a bibliometric analysis to investigate thought the knwoledge of scientific literature of all publications citing the RF article with the science mapping approach. As defined by Tijssen and van Raan [14], science mapping plays a crucial role in the study of knowledge structures underlying research and development (R&D) developments. It depicts the structural and dynamic aspects of a scientific research domain [8] from a quantitative and qualitative viewpoint.

As described by Chen [7], the unit of analysis in science mapping is a domain of scientific knowledge that is reflected through an aggregated collection of intellectual contributions from members of a scientific community or more precisely defined specialties.

The review process was performed using bibliometrix, an R-package (`http://www.bibliometrix.org`) that provides a set of tools for quantitative research in bibliometrics and scientometrics [1]. Bibliometrix is a unique tool according to a logical bibliometric workflow and incorporates a wide variety of different analyses.

3 Data collection and findings

To retrieval the bibliometric data, we queried the Web of Science (WoS) indexing database on October 2021. WoS was launched by the Institute for Scientific Information (ISI) and now it is maintained by Clarivate Analytics. It represent of the main databases allowing to explore the literature of several scientific domains. It includes several citation databases specialised on specific fields covering more than 20000 journals, conference proceedings and books [2].

This systematic literature review and meta-analysis is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines, illustrating the outcomes of the literature searches and article selection process [11].

We downloaded 43.887 publications citing the RF article from WoS. We filtered our collection by selecting only the publications classified as articles and reviews from January 2001 to October 2021. Moreover, we retrieved only the publications written in English. The final collection included 34.713 publications.

In Table 1 there are some descriptive information about the whole collection.The publications has been published on 5.491 sources and collected over one million of references. The collection is charaterized by only the 3.25 % of reviews. The number of single-authored documents is 911. This means that most of publications were written by at least two authors. Indeed the average number of authors per document is 3.

The Figure 2 shows the annual scientific production which provides an overview of the number of papers that have cited RF in the last 20 years. The annual growth rate is on average 49.6%. To date, the number of publications is almost equal to the number of publications published in the 2020. This highlights the constant need for researchers to use this methodology in their works.

In Figure 3 a word cloud of the most frequent is reported. To perform this analysis, we considered the KeyWords Plus (ID) used in the different documents. The ID are words or phrases that frequently appear in the titles of an article's references but do not appear in the title of the publication itself. Their generation is based upon a special algorithm [9] that is unique to WoS databases. The most frequent ID is "classification". This means that the RF is used in many classifications works such as text classification and image classification. [13].

MAIN INFORMATION ABOUT DATA	
Timespan	2001:2021
Sources (Journals, Books, etc)	5491
Documents	34713
Average years from publication	3.38
Average citations per documents	21.88
Average citations per year per doc	4.082
References	1066621
DOCUMENT TYPES	
Article	33586
Review	1127
DOCUMENT CONTENTS	
Keywords Plus (ID)	39276
Author's Keywords (DE)	63024
AUTHORS	
Authors	109632
Author Appearances	200304
Authors of single-authored documents	911
Authors of multi-authored documents	108721
AUTHORS COLLABORATION	
Single-authored documents	1092
Documents per Author	0.317
Authors per Document	3.16
Co-Authors per Documents	5.77
Collaboration Index	3.23

Fig. 1 Main information about data

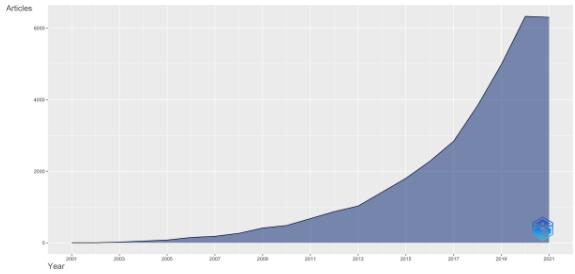


Fig. 2 Annual Scientific Production - Timespan 2001:2021

4 Conclusion

The RF is a user-friendly, intuitive, and very fast approach. It less training time than Decision Tree and Support Vector Machine. RF becomes functional in cases of large datasets because it helps avoid the problem of overfitting by managing the noises present in datasets. As seen in our work, thanks to these capabilities, RF, proposed by Breiman [4], is gaining more and more popularity in the research community for classification and prediction tasks, even twenty years after its publication.

Future developments will be devoted on a detailed analysis of the massive collection used in this work through quantitative and statistical advanced methods. In

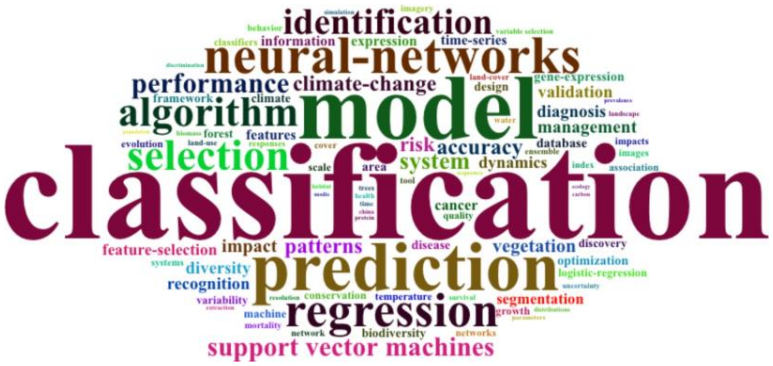


Fig. 3 Word cloud of the most frequent ID

particular, the research domains that use the RF method will be identified to understand what is its main use in different research areas.

References

1. Aria, M., Cuccurullo, C.: bibliometrix: An r-tool for comprehensive science mapping analysis. Journal of informetrics, **11**,(4), pp. 959–975 (2017)
2. Aria, M., Misuraca, M., Spano, M.: Mapping the evolution of social research and data science on 30 years of social indicators research. Social Indicators Research, **149**,(3), pp. 803–831 (2020)
3. Breiman, L.: Bagging predictors. Machine learning, **24**, (2),pp. 123–140 (1996)
4. Breiman, L.: Random forests. Machine learning, **45**, (1), pp. 5–32 (2001)
5. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and regression trees. Routledge (2017)
6. Broadus, R. N.: Toward a definition of "bibliometrics". Scientometrics, **12**, (5-6), pp. 373–379 (1987)
7. Chen, C.: Science mapping: a systematic review of the literature. Journal of data and information science, **2**, (2) (2017)
8. Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F.: Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for information Science and Technology, **62**, (7), pp. 1382–1402 (2011)
9. Garfield, E.: Current eamments. Curr. Contents, **32**, pp. 3–7 (1990)
10. Mitchell, T.: Machine Learning. McGraw Hill (1997)
11. Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. **62** (10), pp. e1-e34, (2009)
12. Pritchard, A.: Statistical bibliography or bibliometrics. Journal of documentation, **25**(4), pp. 348–349 (1969)

13. Song, Q., Liu, X., Yang, L.: The random forest classifier applied in droplet fingerprint recognition. In 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) pp. 722-726, IEEE, (2015)
14. Tijssen, R. J., Van Raan, A. F.: Mapping changes in science and technology: Bibliometric co-occurrence analysis of the R&D literature. Evaluation Review, **18**,(1), pp. 98–115 (1994)
15. Van Raan, A. F.: Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics, **62**,(1), pp. 133–143 (2005)

A study on the GEV activation function for classification of class imbalance data

Hyebin Park, Juyoung Hong[1]

Abstract The classification problems for the imbalance data occur frequently in our lives, and it is important to solve them well. Therefore, we propose a method combining the generalized extreme value (GEV) activation function and the cost-sensitive learning method and over-sampling in a simple neural network model. In order to check the performance of the proposed method, 100 data sets were employed and 5 evaluation metrics were considered. The one-way analysis of variance (ANOVA) and post-hoc tests were performed under the 5% significance level. 162 out of 500 combinations of data sets and 5 evaluation metrics, that is, 32.4% of total showed a significant difference under the 5% significance level. The optimal sampling ratio is judged to be 20:1 and when we compared the results of proposed method and SOTA model, excellent results were obtained in all five data sets.

Key words: Activation function, Class imbalance, Over-sampling, Sigmoid function

1 Introduction

Nowadays, the development of Internet technologies such as social network service (SNS) and the Internet of things (IOT) has enabled us to collect a lot of data. Various types of data are being generated rapidly, and the amount of data is also increasing. Statistical models using such big data help us to make quick and accurate decisions. For this reason, machine learning has become an essential tool in all industries to understand data and improve productivity.

[1] Hyebin Park, Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea; email: central__@naver.com

Juyoung Hong, Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea; email: hjy_stat@naver.com

Among them, classification refers to classifying each instance into a given class by deriving a meaningful relationship between input variable and target variable. Classification problem is a very important problem that occurs very often. Traditional classification algorithms assume that the number of samples between classes is approximately equal. But in reality, that is rarely the case. Such a case in which a specific class appears more frequently than other classes is said to be a class imbalance problem, and it exists in real life such as medical diagnosis, fire detection and fraudulent transaction detection.

Previously, this problem was solved through re-sampling method or cost-sensitive learning method, but recently, there are some trials using the GEV activation function to solve class imbalance problem. Wang et al.[1] used GEV as the link function of GLM, and Lkhagvadori et al.[2] improved classification performance by using a neural network model that has Gumbel distribution as an activation function. Most recently, J. Bridge et al.[3] used GEV activation function in a convolution neural network (CNN) model that diagnoses COVID-19.

2 Methods

When solving a classification problem using a multi-layer perceptron, we often use sigmoid as activation function. The sigmoid function is calculated as in Equation 1 and has a symmetrical structure as shown in Figure 1.

$$sigmoid(x) = \frac{1}{1 + e^{x}} \quad \cdots \textbf{ Equation 1}$$

Our proposed method is to use the CDF of the GEV distribution as the activation function instead of the sigmoid function, because it makes all real inputs to a value between 0 and 1. The GEV distribution has three parameters such as $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}$, and in this study, each parameters were estimated using back propagation method with the weights of the neural network model. The GEV activation function is calculated as in Equation 2, and has an asymmetric structure as shown in Figure 2.

$$G(s) = exp\{-[1 + \xi s]^{-1/\xi}\},\ s = \frac{x-\mu}{\sigma} \quad \cdots \textbf{ Equation 2}$$

Defined on $\{s : 1 + \xi s > 0\}$

Where $-\infty < s < \infty, -\infty < \mu < \infty, \sigma > 0, -\infty < \xi < \infty$

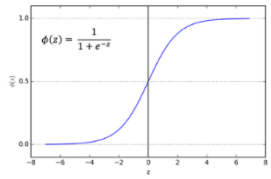


Figure 1: Sigmoid function

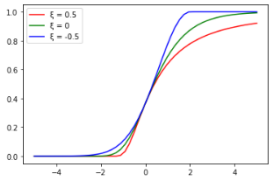


Figure 2: Cumulative distribution function of GEVD

To compare the performance of the proposed method, we considered the following 5 cases for 100 KEEL imbalanced data sets.

1. (Baseline) MLP using sigmoid activation
2. MLP using GEV activation function
3. MLP using GEV activation and Thresholding
4. MLP using GEV activation, Thresholding and Focal Loss
5. MLP using GEV activation, Thresholding, Focal Loss and Over-Sampling

The data used in this experiment are shown in Table 1. The asymmetry ratio was calculated by dividing the number of majority class samples by the number of minority class samples, larger this value means the more severe asymmetry.

Table 1: The data used in this experiment (5 out of 100 KEEL imbalance data sets)

Data name	*# of samples*	*# of input variables*	*Imbalance ratio*
abalone19	4,174	8	129.44
abalone20_vs_8-9-10	1,916	8	72.69
kr-vs-k-zero_vs_fifteen	2,193	19	80.22
pocker-8_vs_6	1,477	9	85.88
pocker-8-9_vs_5	2,075	9	82.00

For a more reliable result, the average of the results obtained by changing the seed (30 times) was compared, and for each data, 5 evaluation indicators [4] (Equation3) suitable for unbalanced data were evaluated. The structure of the neural network model used in the experiment is shown in Figure 3, and the 5 evaluation indicators are shown in Equation 3. For the first four indicators, the higher the value, the better, and the last one, the lower the value, the better. For the reliability of comparison, all hyper parameters such as batch size were made the same. We used one-way ANOVA and post-hoc tests using the results.

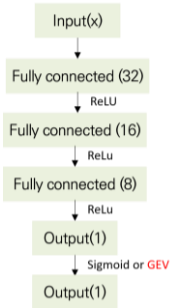


Figure 3: Structure of neural network for this study

F1-score = $\frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}$

Geometric-Mean = $\sqrt{\text{TPR(Recall)} \times \text{TNR(Specificity)}}$

Balanced Accuracy = $\frac{1}{2} \times (\text{TPR} + \text{TNR})$ ⋯ **Equation 3**

Area Under the ROC Curve (AUC)

Brier Inaccuracy = $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{1}\left(\hat{p}(c = j, x^i) - p(c = j, x^i)\right)^2$

3 Results

A summary of the experimental results is shown in Table 2. If there's significant differences of metrics under 5%, we counted the number of better results when we compared method 1 and 5. Better values are marked in red for better visibility. It's never been nice if the GEV activation function was used alone. In particular, it is interesting to note that as you move down the table, the results look better.

Table 2: Example of experiment result (Data : abalone19)

Data	*method*	*F1-score*	*Geometric-Mean*	*Area Under Curve*	*Balanced Accuracy*	*Brier Inaccuracy*
abalone19	(1)	0.0 (0.0)	0.0 (0.0)	0.794 (0.084)	0.5 (0.0)	0.016 (0.0)
	(2)	0.0 (0.0)	0.0 (0.0)	0.659 (0.143)	0.5 (0.0)	0.019 (0.013)
	(3)	0.033 (0.021)	0.662 (0.16)	0.659 (0.143)	0.687 (0.105)	0.019 (0.013)
	(4)	0.045 (0.023)	0.733 (0.165)	0.760 (0.119)	0.757 (0.095)	0.037 (0.012)
	(5)	0.044 (0.015)	0.762 (0.057)	0.781 (0.068)	0.770 (0.056)	0.039 (0.008)

Table 3: Comparing with SOTA model (GEV-NN)

Data	*Geometric-Mean (Proposed)*	*Geometric-Mean (GEV-NN)*	*Area Under the ROC Curve (Proposed)*	*Area Under the ROC Curve (GEV-NN)*
abalone19	0.762	0.7247	0.781	0.7419
abalone20_vs_8-9-10	0.908	0.884	0.935	0.9009
kr-vs-k-zero_vs_fifteen	1	1	1	1
pocker-8_vs_6	0.998	0.9714	0.999	0.966
pocker-8-9_vs_5	0.752	0.5165	0.719	0.408

4 Summary and Discussion

In this experiment, a toy model experiment using the KEEL imbalance dataset was conducted, but it needs to be applied to real-world data set such as rainfall, financial data, etc. Since the superiority of the model is shown differently depending on the evaluation index, a comparison method that considers the characteristics of the model and data is needed. Better results can be expected if the two hyperparameters for Focal Loss are adjusted. The optimal ratio of Over-Sampling is about 20:1, but it is difficult to apply OS depending on the sample size or imbalance ratio.

In addition, it is currently applied only to the case of binary classification, but it can be extended to multi-class classification in the future.

Acknowledgments

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No.2020R1I1A3069260) and BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

1. Wang, X., Dey, D.K.: Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. Ann. Appl. Stat. **4(4)**, 2000--2023 (2010) doi: 10.1214/10-AOAS354
2. Munkhdalai, L., Munkhdalai, T., Ryu, K.H.: GEV-NN: A deep neural network architecture for class imbalance problem in binary classification. Knowledge-Based Systems. **194**, 105534 (2020) doi: 10.1016/j.knosys.2020.105534
3. J. Bridge, et al.: Introducing the GEV Activation Function for Highly Unbalanced Data to Develop COVID-19 Diagnostic Models. IEEE Journal of Biomedical and Health Informatics. **24(10)**, 2776--2786 (2020) doi: 10.1109/JBHI.2020.3012383
4. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data. **6(27)**, (2019) doi: 10.1186/s40537-019-0192-5

Session of free contributes SCL7 – *Society and Disparity*
Chair: Pasquale Sarnacchiaro

The gender gap in lifespan disparity as a social indicator of international countries: A fuzzy clustering approach

Il divario di genere nella longevità come indicatore sociale per paesi internazionali: Una classificazione fuzzy

Leonardo Salvatore Alaimo, Pierpaolo D'Urso and Andrea Nigri

Abstract In this paper, we apply a Dynamic Time Warping-based Fuzzy *C*-Medoids clustering model with Exponential transformation (DTW-Exp-FCMd) to the multivariate time series of the gender gap in longevity indicators, namely life expectancy and lifespan disparity at birth. Data are collected from the Human Mortality Database, obtaining longevity measures using mortality rates, and then computing the gap as the ratio F/M. By exploiting the clustering information we are able to gain better insights into the gender gap in longevity over time and countries. Its monitoring might play a crucial role for the public and private sector, and in order to gain long-term sustainability goals.

Abstract *In questo articolo, applichiamo un modello di classificazione fuzzy, il Dynamic Time Warping-based Fuzzy C-Medoids clustering model with Exponential transformation (DTW-Exp-FCMd), alle serie temporali multivariate del divario di genere in due indicatori di longevità, l'aspettativa di vita e la disuguaglianza di durata della vita alla nascita. I dati sono raccolti dallo Human Mortality Database e le misure di longevità sono ottenute utilizzando i tassi di mortalità, e poi calcolando il gap come rapporto F/M. Sfruttando le informazioni ottenute dalla classificazione, siamo in grado di ottenere una migliore comprensione del divario di genere nella longevità nel tempo e nei paesi. Il suo monitoraggio potrebbe giocare un ruolo cruciale per il settore pubblico e privato, e per ottenere obiettivi di sostenibilità a lungo termine.*

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

Pierpaolo D'Urso
Department of Social Sciences and Economics, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy, e-mail: pierpaolo.durso@uniroma1.it

Andrea Nigri
Department of Social and Political Sciences, Bocconi University, Milan, Italy, e-mail: andrea.nigri@unibocconi.it

Key words: Lifespan disparity, Life expectancy, Time series, Fuzzy clustering, Dynamic Time Warping-based Fuzzy *C*-Medoids clustering model with Exponential transformation

1 Introduction

Health indicators as life expectancy at birth are crucial in measuring the quality of life ([2];[3]). Its appeal relies on the ability to enclose and summarise all the factors affecting longevity. Indeed, even though developed countries have recently experienced stability in mortality and long periods of increasing life expectancy, diverging trends emerged, with some countries stagnating, some decelerating, showing different speeds [6]; [5] It is important to note that populations with the same level of life expectancy may have different variability in the distribution of the age at death. An additional indicator able to provide a measure of dispersion in the age-at-death is the lifespan disparity, which can be considered an estimation of the heterogeneity at the population level ([1]). Albeit this aspect is still neglected in the study of well-being and social indicators at the aggregate level, it might be crucial in the economics, public system, as well as in determining sustainability goals. The study plays a major role when considering gender differences in longevity heterogeneity around the globe. Therefore, we strongly believe that it is urgent to delve into the subject as, so far, there are no scientific contributions that provide monitoring of the gender gap in life span disparity, even more so from a global perspective. Moving from these considerations, we analyse the patterns of both lifespan disparity and life expectancy gender gap series, by identifying different phases and transitions that allow clustering countries according to their longevity dynamics. Indeed, changes in health indicators for a group of countries might be attributable to common factors, such as similar socioeconomic circumstances shared improvements in public health, and medical technology, then it makes sense to treat them jointly.

The paper is organised 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.

2 Data description and methods

2.1 Demographic measures

Period life expectancy at birth is the most widely used indicator of population health and longevity. It refers to the expected average age at death for a synthetic cohort of newborns, that experience the mortality risks of that time throughout their lifespan. We define the life expectancy at age x and time t in a given population, as follows:

$$e_{x,t} = \frac{\int_x^{\infty} S(y,t)dy}{S(x,t)} \tag{1}$$

Where, where $S(x,t) = \exp\left(-\int_0^x \mu(a,t)da\right)$ and $\mu(a,t)$ are the survival function and the force of mortality respectively.

Thus, we can introduce the lifespan disparity as an indicator representing the life expectancy lost due to death by an individual aged x at time t.

Formally the lifespan disparity at birth is defined as follows:

$$e^{\dagger}_{0,t} = -\int_0^{\infty} S(a,t) \cdot \ln S(a,t)da \tag{2}$$

We consider the gender gap in lifespan disparity ($G^{(e^{\dagger}_0)} = e^{\dagger}_{0,t,Female}/e^{\dagger}_{0,t,Male}$) and in life expectancy ($G^{(e_0)} = e_{0,t,Female}/e_{0,t,Male}$) at birth, using historical data provided by the Human Mortality Database (HMD).

2.2 Clustering method

In this work, we deal with a three-way time data array $\mathbf{X}$ of the type "units $\times$ variables $\times$ times" [3]. Specifically, we consider 33 counties and the two gender gap variables, $G^{(e^{\dagger}_0)}$ and $G^{(e_0)}$. Data are available from 1990 to 2015. Formally: $\mathbf{X} \equiv \{x_{ijt} : i = 1,\ldots,33;\ j = G^{(e^{\dagger}_0)}, G^{(e_0)};\ t = 1990,\ldots,2015\}$. We must clarify that observed levels equal to 1 indicate an absence of a gender gap for both variables. On the contrary, values different (higher or lower) from 1 represent clear evidence of diverging behaviours in longevity between male and female populations. It is worth noting that life expectancy has a growing dynamic; thus, high values correspond to improvements in longevity. On the contrary, the measure of lifespan disparity is characterised by a monotonous decreasing trend, where lower values represent less dispersion and therefore, improvements in longevity levels.

The clustering method chosen is the Dynamic Time Warping-based Fuzzy C-Medoids clustering model with Exponential transformation (DTW-Exp-FCMd) (for details, see: [3]), based on the Exponential transformation of the Dynamic Time warping distance [8] and well-known and used in literature (for instance, see: [2]):

$$\begin{cases} \min: & \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\, _{exp}d^2_{DTW}(\mathbf{X}_i, \widetilde{\mathbf{X}}_c) = \\ & \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\left[1-\exp\left\{-\beta\, d^2_{DTW}(\mathbf{X}_i, \widetilde{\mathbf{X}}_c)\right\}\right] \\ s.t. & \sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0 \end{cases} \tag{3}$$

where

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C}\left[\frac{\left[1-\exp\left\{-\beta\, d^2_{DTW}(\mathbf{X}_i, \widetilde{\mathbf{X}}_c)\right\}\right]}{\left[1-\exp\left\{-\beta\, d^2_{DTW}(\mathbf{X}_i, \widetilde{\mathbf{X}}_{c'})\right\}\right]}\right]^{\frac{1}{m-1}}}. \tag{4}$$

The optimal partition, C, has been chosen by means of the the Xie-Beni criterion [7]:

$$\min_{C \in \Omega_C} : I_{XB} = \frac{\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ik}^{m} d_{DTW}^{2}(\mathbf{X}_i, \widetilde{\mathbf{X}}_c)}{I \min_{c,c'} d_{DTW}^{2}(\widetilde{\mathbf{X}}_c, \widetilde{\mathbf{X}}_{c'})} \tag{5}$$

where Ω_C represents the set of possible values of $C\,(C < I)$. The optimal number of clusters C is identified in correspondence with the lower value of I_XB (for more detail, please see: [2]). The used time series clustering approach presents all the general advantages connected to the fuzzy theory in a clustering framework, is sensitive in capturing the dynamic characteristics of the time series and inherits the advantage connected to DTW distance (for details, please see: [2]).

3 Application and results

We compute the Xie-Beni index for $2 \leq c \leq 6$ obtaining these results: $I_{XB}(c=2) = 0.079$; $I_{XB}(c=3) = 0.061$; $I_{XB}(c=4) = 0.143$; $I_{XB}(c=5) = 0.146$; $I_{XB}(c=6) = 0.331$. Accordingly, we selected the 3 clusters partition. For the evaluation of the fuzziness, we need to specify a cut-off point for the membership degree. Given a three-cluster situation, if the membership degrees in two clusters are between 0.3 and 0.7, we can affirm that the situation of the corresponding unit is fuzzy. Consequently, the value 0.7 has been chosen as cut-off. Therefore, those countries that do not have at least that value as membership degree to a cluster are considered fuzzy (for more information on the choice of cut-off, please see: [2, 4]).

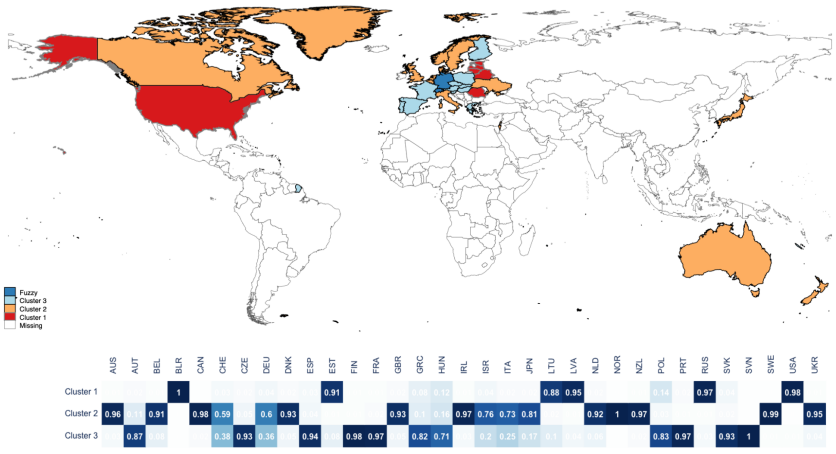


Fig. 1: Clusters composition. 33 Countries. Years 2010–2015

The optimal solution identifies 3 Clusters: Cluster 1 (6 countries) with Belarus (BEL) as medoid, Cluster 2 (14 countries) represented by Norway (NOR) and Cluster 3 (11 countries) with Slovenia (SVN) as medoid. There are two fuzzy countries, Germany (DEU) and Switzerland (CHE). Figure 1 shows the subdivision of the countries according to the cluster to which they belong and the matrix with the membership degrees.

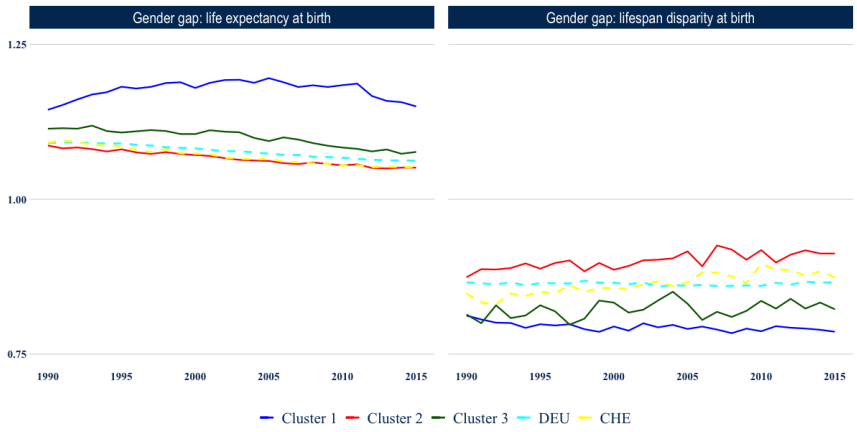


Fig. 2: Comparison among Cluster 1, Cluster 2, Cluster 3 and fuzzy counties: Switzerland (CHE) and Germany (DEU).

The membership of each country to its respective cluster, except for the fuzzy counties, is clear and unambiguous. Clusters are clearly characterised, as we can observe in Figure 2. Countries of Cluster 1 present the worst situation in terms of gender equality. It is interesting to note how it includes the countries of the former Soviet bloc, including Russia (RUS), and the USA. Cluster 3 includes those countries having better situation in terms of gender equality, both for life expectancy and lifespan disparity. Finally, Cluster 2 consists of countries with a intermediate level of gender equality (between Cluster 1 and Cluster 3). CHE and DEU present a fuzzy behaviour, staying between Cluster 3 and Cluster 2. We can affirm that they are countries with an intermediate level and a propensity to increase it.

4 Conclusions

The results of this preliminary study show that the evolution of the gender gap in both longevity indicators follows a homogeneous pattern with different timing and transitions. The identified clusters reflect the demographic theoretical background of life expectancy and lifespan disparity at birth phases and transitions. Our findings

support the exists of a best practice cluster characterised by lower levels of gaps, composed of countries that recently achieved constant growth in life expectancy levels. Furthermore, we bring evidence of the worst scenario, where the gender gap monitoring might play a crucial role for the public and private sector, pension schemes, and health policies are prime examples. In this context, our analysis displays those countries that experienced stagnation (USA) or deceleration in longevity (Eastern countries). Here can be found the Russian case, where the literature well-documents the gender difference in life expectancy due to alcohol consumption and the excess of cardiovascular mortality for the male population.

References

1. Aburto, J.M., van Raalte A.: Lifespan Dispersion in Times of Life Expectancy Fluctuation: The Case of Central and Eastern Europe. Demography **55(6)**, 2071–2096 (2018) doi: 10.1007/s13524-018-0729-9.
2. D'Urso, P., Alaimo, L.S., De Giovanni, L., Massari, R.: Well-Being in the Italian Regions Over Time. Social Indicators Research (2020) doi: 10.1007/s11205-020-02384-x.
3. D'Urso, P., De Giovanni, L., Massari, R.: Using Dynamic Time Warping to Find Patterns in Time Series. International Journal of Approximate Reasoning **99**, 12–38 (2018).
4. Maharaj, E.A., D'Urso, P.: Fuzzy Clustering of Time Series in the Frequency Domain. Information Sciences **181(7)**, 1187–1211 (2011).
5. Nigri, A., Levantesi, S., Marino, M.: Life expectancy and lifespan disparity forecasting: a long short-term memory approach. Scandinavian Actuarial Journal **2**, 110-133 (2021). doi: 10.1080/03461238.2020.1814855.
6. Nigri, A., Barbi, E., Levantesi, S.: The relationship between longevity and lifespan variation. Statistical Methods and Applications (2021). doi: 10.1007/s10260-021-00584-4.
7. Xie, X.L., Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis & Machine Intelligence, **13(8)**, 841–847 (1991).
8. Velichko, V.M., Zagoruyko, N.G.: Automatic Recognition of 200 Words. International Journal of Man-Machine Studies **2(3)**, 223–234 (1970).

Modelling scale effects via a Bayesian approach: an application to decision making in public sector

Modellazione degli effetti di scala tramite un approccio bayesiano: un'applicazione al processo decisionale nel settore pubblico

Maria Iannario and Claudia Tarantola

Abstract We present a Bayesian approach for the analysis of rating data when a scaling component is taken into account. Model-based probability effect measures for comparing distributions of several groups, adjusted for explanatory variables affecting both location and scale components, are computed. Markov Chain Monte Carlo techniques are implemented to obtain parameter estimates and the mentioned measures. An analysis on students' evaluation of a university orientation service is carried out to assess the performance of the method and make more valuable the decision making process of university players (stakeholders).

Abstract *Il contributo presenta un approccio bayesiano per l'analisi dei dati rating quando una componente di variabilità nei dati è presente. Sono calcolate misure di probabilità basate su modelli per confrontare le distribuzioni di diversi gruppi, condizionate alle variabili esplicative che influenzano sia la componente di posizione che di variabilità del tratto latente. Tecniche Markov Chain Monte Carlo sono implementate per ottenere le stime dei parametri e le misure menzionate. Al fine di valutare la proposta metodologica e contributire al processo decisionale degli attori universitari (stakeholder), viene riportata un'analisi sulla valutazione da parte degli studenti di un servizio di orientamento universitario.*

Key words: Heterogeneity of variances, MCMC, ordinal responses, ordinal superiority measures, scale effects.

Maria Iannario
Department of Political Sciences, University of Naples Federico II, e-mail: maria.iannario@unina.it

Claudia Tarantola
Department of Economics and Management, University of Pavia, e-mail: claudia.tarantola@unipv.it

Maria Iannario and Claudia Tarantola

1 Introduction

Rating surveys to assess the quality of services are becoming one of the main tools to measure respondents' satisfaction with respect to the service received and are representing a widespread practice to evaluate the performance in public sectors (see [8], among others, and reference therein). University curriculum counselor represents one of the recent area where students' evaluation has become a landmark for improving services. We focus on this issue analysing a set of data resulting from a survey conducted in 2002 at University of Naples Federico II. The interviewed were asked to reply to some questions regarding the orientation services located in the 13 Faculties. In Section 3 we provide a brief description of the data; further details are in [5].

To assess the service several scores on rating scales are reported by students. Cumulative models with proportional assumption ([1]) are the main candidates to analyse them. However, if unobserved heterogeneity related to the latent variable behind the observed score is present, scale effects in the regression structure with ordinal responses are needed. Hence, we relay on the location-scale models ([7]), extended in the Bayesian framework ([3, 6]). The latter allows us to improve the flexibility in specifying the model and enhances the accuracy in providing parameter estimates. The implementation of the model, the probability-based measures for comparing clusters on ratings, and marginal effects to address the interpretation of the results on the extreme categories of the rating scales are the focus of the contribution. Results favor the proposed models and show significant subject heterogeneity at baseline.

2 Model description

We consider the following setting. Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ be a sample generated by an ordinal random variable $Y \sim G(y)$ on the support $\{1, \ldots, k\}$, where k is a known integer. We indicate with Y_i^* the underlying (continuous) latent variable such that when $\alpha_{j-1} < Y_i^* \leq \alpha_j$, then $Y_i = j, j = 1, 2, \ldots, k$. Here $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_k = +\infty$ are the thresholds of Y^*.

In our context, Y_i is the rating expressed by the i-th student on a specific question concerning the evaluation of the university orientation service. For each student, we collect information $\mathscr{I}_i = (y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, where y_i is the observed value of the rating and $\boldsymbol{x}_i$ is a row vector of the matrix $\boldsymbol{X}$ which includes all the relevant covariates useful for characterising students' profile.

When $p \geq 1$ covariates are relevant for explaining Y^*, the latent regression model behind the process of response is $Y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \sigma_i\varepsilon_i$, $i = 1, 2, \ldots, n$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the covariates coefficients. In the latent regression σ_i is the standard deviation of the noise variable $\varepsilon \sim F_\varepsilon(.)$, which may depend on covariates yielding $\sigma_i = \exp(\boldsymbol{z}_i\boldsymbol{\gamma})$. Here $\boldsymbol{z}_i$ is a row vector of the matrix $\boldsymbol{Z}$ which includes all the $q \geq 1$ relevant covariates and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ are the related covariates coefficients.

Then, the probability mass function of Y_i, for $j = 1, 2, \ldots, k$, is

$$\begin{aligned} Pr(Y_i = j \mid \boldsymbol{\theta}, \boldsymbol{x}) &= Pr\left(\alpha_{j-1} < Y_i^* \leq \alpha_j\right) \\ &= F_\varepsilon\Big[(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma_i\Big] - F_\varepsilon\Big[(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma_i\Big]. \end{aligned}$$

Among the alternative choices for $F_\varepsilon(.)$ we focus on the logit link function for easiness of interpretation and robustness properties. Since we do not have relevant prior information, we use non informative priors on all parameters of interest, letting the data guide the behaviour of the posterior distributions. Finally, we rely on MCMC methods to obtain posterior samples.

2.1 Ordinal superiority measures

Ordinal superiority measures for group comparison presented by [2], and implemented in [4], has been here extended to deal with the scaling effect in a Bayesian context. Given a dichotomous variable z_{ik} playing a role in the scale parameter we use these measures to compare the probability that an observation from one group g_0 (i.e., not-frequently users of the service) is scored above an independent observation from the alternative group g_1 (i.e., frequently users of the service).

Let indicate with $\boldsymbol{x}_{\setminus d}$ the set of all covariates with the exception of z_{ik}. At a generic iteration t ($t = 1, \ldots T$) and for a specific value $\boldsymbol{x}^*_{\setminus d}$ the ordinal superiority measure Δ is given by

$$\begin{aligned} \Delta^t(\boldsymbol{x}^*_{\setminus d}) &= Pr^t(Y_{g_0} > Y_{g_1}) - Pr^t(Y_{g_1} > Y_{g_0}) \\ &= \sum_{l>k} \hat{\pi}^t_{0l}(\boldsymbol{x}^*_{\setminus d})\hat{\pi}^t_{1k}(\boldsymbol{x}^*_{\setminus d}) - \sum_{k>l} \hat{\pi}^t_{0l}(\boldsymbol{x}^*_{\setminus d})\hat{\pi}^t_{1k}(\boldsymbol{x}^*_{\setminus d}) \end{aligned}$$

where $\hat{\pi}^t_{0r}(\boldsymbol{x}^*_{\setminus d}) = \hat{Pr}(Y = j; d = 0, \boldsymbol{x}^*_{\setminus d})$ is the fitted value obtained from the examined model for g_0; $\hat{\pi}^t_{1r}(\boldsymbol{x}^*_{\setminus d})$ is obtained in a similar way for g_1.

We use the MCMC output to obtain the posterior estimates of Δ as follows

$$\widehat{\Delta(\boldsymbol{x}^*_{\setminus d})} = \frac{1}{T}\sum_t \Delta^t(\boldsymbol{x}^*_{\setminus d}).$$

A value of $\widehat{\Delta(\boldsymbol{x}^*_{\setminus d})}$ greater than zero indicates that it is more likely to obtain a higher rating in g_0 than in g_1. Alternatively one can calculate the γ measure having null value equal to 0.5. Its Bayesian estimate is given by

$$\widehat{\gamma(\boldsymbol{x}^*_{\setminus d})} = \frac{1}{T}\sum_t \gamma^t(\boldsymbol{x}^*_{\setminus d}) \text{ with } \gamma^t(\boldsymbol{x}^*_{\setminus d}) = (\Delta^t(\boldsymbol{x}^*_{\setminus d}) - 1)/2.$$

3 University curriculum counselor data analysis

Data consist of 2179 observations on 12 variables. Respondents were asked to express on a seven-points scale how they evaluate university orientation service with respect to the Office hours (*Office*) - number of hours when the offices are open to the students. From a preliminary analysis of the data we identified the following relevant covariates: *Gender* of the respondent, a dichotomous variable (1=female, 0=male); *Age*, a continuous variable (17 to 51, mean= 22.6); *Freq serv*, a factor with levels: 0 = for not-frequently users, 1 = for frequently users; and a factor variable related to *Area* of study for the different Faculties: 0=Scientific, 1=Health Science and 2=Humanistic.

The Bayesian estimates of the location and scale parameters are reported in Table 1 (posterior mean, MCMC Standard Error and 95% credible intervals). These results are obtained via the R package `brms` (Bayesian regression model using "Stan"); see [3]. Standard convergence diagnostics have been considered. The estimated thresholds are $\hat{\alpha}_1 = -2.46$*(0.31)*, $\hat{\alpha}_2 = -1.75$*(0.30)*, $\hat{\alpha}_3 = -0.91$*(0.29)*, $\hat{\alpha}_4 = -0.14$ *(0.29)*, $\hat{\alpha}_5 = 1.15$ *(0.29)* and $\hat{\alpha}_6 = 2,38$*(0.30)*. We run in parrel 4 chains of 2000 iteration with a burnin period of 1000 iteration each. The Bayesian estimate of the standard deviation is obtained from the posterior samples of log-disc (log-discrimination) with disc corresponding to the inverse of the standard deviation.

Table 1 Bayesian estimates for the location-scale model

	Estimate	Sd	l-95% CI	u-95% CI
Age	0.04	0.01	0.02	0.06
Gender	0.15	0.08	-0.01	0.32
Freq serv	0.70	0.11	0.49	0.92
Health Science	0.65	0.12	0.42	0.89
Humanistic	-0.64	0.12	-0.87	-0.41
log_disc_Freqserv	-0.32	0.05	-0.42	-0.22
sd_disc_Freqserv	1.38	0.07	1.25	1.52

In order to evaluate the presence of heterogeneity in our data we rely on conditional effects on specific covariates with respect to the response variable. First of all we concentrate upon the variable *Freq serv* affecting both the location and the scale component. In the upper panel of Figure 1 we provide a visual representation of the estimated relationship between *Freq serv* and *Office*. The upper panel of figure displays the estimated probabilities of the seven response categories for the two groups. We notice that a higher evaluation is provided by students using more the service. The latter represent also the most heterogeneous group ($\hat{\sigma}_{Freq\ serv} = 1.38$). In the bottom panel of Figure 1 we report the evaluation for the different *Area* of study. Health Science students tend to provide a higher evaluation than Scientific and Humanistic ones. Taking into consideration the variable *Age*, we observe that a higher evaluation of the service is provided by older people whereas for *Gender*, male seems to be more critical (see Table 1).

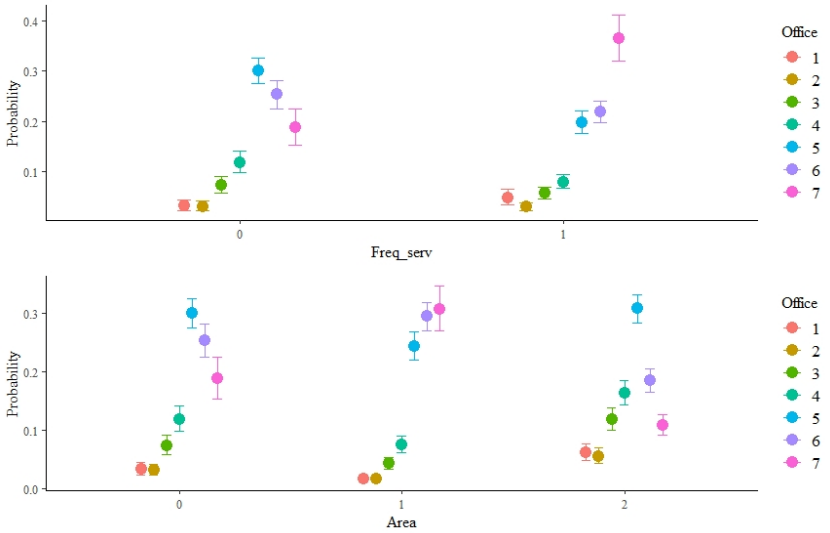


Fig. 1 Marginal effects of *Freq serv* (first panel) and *Area* (second panel) on *Office* evaluation. Points indicate the posterior mean estimates and error bars corresponds to the 95% Credible Intervals.

In Table 2 we report the ordinal superiority measures for *Freq serv* confirming previous results. For example, the negative value of $\widehat{\Delta}$ indicates that there is a higher probability to obtain a higher evaluation in the group g_1 (students who frequently use the service).

Table 2 Bayesian ordinal superiority measures

	Estimate	Sd	l-95% CI	u-95% CI
$\widehat{\Delta}$	-0.15	0.03	-0.21	-0.10
$\widehat{\gamma}$	0.42	0.01	0.40	0.45

Further analyses concerning other aspects of the service may improve the institution overall service strategy. Furthermore, an analysis in which we scrutinize the complex relationships between the latent variable on different levels (universities), exploiting a multilevel framework, may allow us to study how group membership is expected to influence data analysis results.

References

1. Agresti, A.: Analysis of Ordinal Categorical Data, 2nd ed. Wiley, Hoboken (2010)

2. Agresti, A., Kateri, M.: Ordinal Probability Effect Measures for Group Comparsons in Multinomial Cumulative Link Models. Biometrics, **73**, 214–219 (2017)
3. Bürkner, P.: brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, **80**(1), 1–28 (2017)
4. Iannario, M., Tarantola, C.: Effect measures for group comparisons in a two-component mixture model: a cyber risk analysis. In: Balzano, S., Porzio, G. C. Salvatore, R. Vistocco, D., Vichi, M. (eds) Studies in Classification, Data Analysis and Knowledge Organization, pp 97–104. Springer Nature, Switzerland (2021)
5. Iannario, M., Piccolo, D., Simone, R.: CUB: A Class of Mixture Models for Ordinal Data. https://CRAN.Rproject.org/package=CUB (2020)
6. Liddell, T. M., Kruschke, J.K.: Analyzing ordinal data with metric models: What could possibly go wrong? Journal of Experimental Social Psychology, **79**, 328–348 (2018)
7. McCullagh, P.: Regression Models for Ordinal Data. Journal of the Royal Statistical Society. Series B, **42**, 109–142 (1980)
8. Poister, T.H., Gary, T.H.: Citizen Ratings of Public and Private Service Quality: A Comparative Perspective. Public Administration Review, **54**, 155–160 (1994)

Board Gender Diversity and Social engagement: evidence from the banking industry

Diversità di genere del board e impegno sociale: evidenze dal settore bancario

Francesco Gangi, Lucia Michela Daniele and Maria Coscia

Abstract
Board gender diversity (BGD) and corporate social responsibility (CSR) are strictly linked since the former pertains to corporate governance (CG) which represents a pillar of CSR. Currently, the banking industry is experiencing higher CSR engagement. Accordingly, from a systemic perspective, the current paper investigates whether BGD is a determinant of better social performance of banks. To address this question, we run a fixed-effects regression analysis based on a sample of 137 international banks. Our findings show that BGD positively impact on the social engagement of banks. This result is consistent with the conflict resolution thesis which routes in the stakeholder theory.

Abstract
La diversità di genere del board (BGD) e la responsabilità sociale d'impresa (CSR) sono strettamente collegate, in quanto la BGD rappresenta un meccanismo efficace di corporate governance (CG). Attualmente, il settore bancario sta sperimentando un maggiore impegno in CSR. Pertanto, adottando una prospettiva sistemica, il presente studio indaga la BGD quale fattore determinante di migliori performance sociali delle banche. Il presente studio adotta una analisi di regressione a effetti fissi basata su un campione di 137 banche internazionali. I risultati mostrano che la BGD impatta positivamente sull'impegno sociale delle banche. Ciò è coerente con la prospettiva della "conflict resolution hypothesis" e la teoria degli stakeholder.

Key words: Gender Diversity, CSR, Social Engagement, Bank Industry, SDGs.

1 Introduction

The 5th Sustainable Development Goals (SDGs), the gender equality or parity, represents a global priority for achieving a more sustainable and resilient world (United Nations, 2017). At the same time, the Global Gender Gap Report 2020 indicated that female economic participation is an important challenge. Globally, the presence of women on corporate boards appears still limited (22%) (WTF, 2020).

[1] Francesco Gangi, Department of Economics, University of Campania, Luigi Vanvitelli; francesco.gangi@unicampania.it

Lucia Michela Daniele, Department of Economics, University of Campania, Luigi Vanvitelli; luciamichela.daniele.gangi@unicampania.it

Maria Coscia, Department of Management Studies and Quantitative Methods, Parthenope University of Naples; maria.coscia001@studenti.uniparthenope.it

Moreover, the Covid-19 has triggered a social crisis for gender equality women's rights (EU, 2021). By investigating the link between CG and CSR, the CSR research stream highlighted that greater presence of women within boards of firms should improve sustainable performance (Naciti, 2019). However, few evidences are currently available on the banking sector (Wu and Shen, 2013). Accordingly, we investigate the impact that BGD on banks' social performance (BSP). Our sample consists of 137 worldwide banks, observed for a time period spanning from 2009 to 2020. Results from the fixed effect panel regressions corroborate the vision of female representation as an important driver to reach greater social performance within bank industry. The remainder of study is structured as follows. Section 2 provides the theoretical background supporting our research question. After the description of the methodology in Section 3, we present the empirical results in Section 4. Finally, Section 5 leads to the discussion of the results and to conclusive remarks.

2 Theoretical Background

Focusing on top management teams, board gender diversity (BGD) constitutes a key corporate governance (CG) mechanism within the wider corporate social responsibility (CSR) (Bear et al., 2010; Harjoto et al., 2015). Indeed, following Renneboog et al (2008), we define CSR as "corporate decisions fostering social, corporate governance, ethical and environmental issues". This definition highlights CG, social responsibility and environmental responsibility as the main pillars of CSR. Accordingly, we analyse the link between BGD and social engagement.

2.1 BGD and social engagement

According to the United Nations, the 5th SDG aims to include women in centre of economies in order to drive more sustainable outcomes. The inclusion of the gender perspective in CSR practices can play a pivotal role in achieving gender equality in the workplace, by providing equal access to opportunities for women on boards (Naciti, 2019). The social inclusion policies are part of the wider CG plans, that are relevant in addressing managers towards the adoption of socially friendly initiatives (Jain and Jamali, 2016; Naciti, 2019). Among the CG mechanisms, BGD may be a determinant for higher CSR commitment (Bear et al., 2010; Harjoto et al., 2015). Regarding the link between BGD and the social engagement, earlier literature (Alazzani et al., 2017; Shakil et al., 2020) highlights the relevance of the female directors on board for promoting activities in line with the social community's expectations. Based on social role theory (Eagly, 1987), women are more sensitive to social needs than their male counterpart (Boulouta,

2013). Specifically, women are more socially-oriented and prone to understand others' needs, then showing greater inclination toward stakeholders claims. As stakeholder relationships are the main drivers of corporate social performance (CSP) (Cosma et al., 2021), this female trait may lead to better relationships with stakeholders and CSP. Furthermore, women bring a healthy mix of knowledge and experience to the board, by improving the quality of the decision making (Lu and Herremans, 2019). As result, since female directors may be particularly sensitive to socially oriented practices (Nielsen and Huse, 2010), BGD can lead to higher BSP. From the stakeholder approach (Freeman, 1984), several reasons support for CSR engagement in the banking industry. First, by implementing socially friendly actions, banks may improve their reputation, as well as, their customers' loyalty (Aramburu and Pescador, 2019). Second, if a bank incorporates sustainability in its lending policy, then it will be less exposed to information risk and adverse selection (e.g., Goss and Roberts, 2011). Third, banks that adopt better CG mechanisms tend to be more socially responsible since this can reduce boycotts risk. Focusing on BGD in banks, Galletta et al. (2021) found that female managers are more socially engaged with stakeholders, than female directors. However, by disentangling the CSR pillars, we can notice that the relationship between BGD and the engagement in socially oriented actions remains still less investigated than other CSR dimensions (e.g., Gangi et al., 2019). Accordingly, we posit the following research question:
RQ: Does BGD positively affects the social engagement of banks?

3 Methodology

The empirical analysis relies on a sample of 147 worldwide banks extracted from Thomson Reuters. We gathered financial indicators from Worldscope database, while for social engagement and board characteristics we adopted the ASSET4 database. We excluded banks for which social scores were not disclosed during the study period. This leads to a panel of 137 banks from 21 countries, for 1555 bank/year observations. The banks' social engagement, as dependent variable, is measured by the Asset4's social pillar score (SOC). As independent variable, we proxy the BGD with the percentage of women on the board of directors. As controls, we adopt other CG mechanisms, such as the percentage of independent directors (B_ind); a dummy variable equal to 1 if banks provide senior executives with compensation mechanisms linked to sustainability targets (SustComp) and 0 otherwise. Additionally, we adopt banks size through the logarithm of total assets (logTA); the level of indebtedness (Leverage) through the ratio between total debt and total equity; the level of coverage through the incidence of loan loss reserve to gross loans (Coverage); the loan-to-deposit ratio (LoanDep); the level of capital expenditure, proxied as the ratio between total expenditures and total assets (Capex). Finally, we consider the gross domestic product per capita (GDPper). To

verify the RQ, we perform a Fixed-Effect (FE) regression analysis. The regression equation is formalized as follows:

$$SOC_{i,t} = \alpha + \beta BoardGD_{i,t-1} + \psi X_{i,t-1} + \varepsilon_i \quad (1)$$

Where SOC_t refers to the social score of bank *i* at time *t*, BoardGD is the BGD measure of bank *i* at time *t-1*, X is a vector of the control variables, and ε is a random error term.

4 Results

Table 1 and 2 provide descriptive statistics and correlation matrix respectively. Table 3 contains the estimates on the relationship between BGD and the social engagement. The findings show that the presence of female directors positively affects the orientation of banks' board of directors towards social issues, at the 5%. This positive association suggests that higher gender equality among the board members increases the level of sensitivity on socially responsible initiatives employed by banks. Focusing on the controls, we find a significant and positive impact of the presence of independent directors and the sustainability compensation policy on the social commitment.

Table 1. *Descriptive Statistics.*

Variables	*Obs*	*Mean*	*Median*	*SD*
SOC	1555	51.22	51.49	25.60
BoardGD	1555	18.87	18.18	13.28
B_ind	1525	61.99	69.23	27.92
SustComp	1560	0.25	0	0.43
logTA	1637	8.66	8.52	1.11
Leverage	1635	63.39	67.81	20.98
Coverage	1470	187.36	159.2	179.26
LoanDep	1615	115.77	95.76	128.38
Capex	1616	1.38	297,598.5	6.37
GDPper	1623	46,792.3	47,099.98	13,417.11

Table 2: *Correlation matrix and VIF.*

Variables	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*VIF*
BoardGD	1.00									1.39
B_ind	0.40	1.00								1.94
SustComp	0.27	0.25	1.00							1.18
logTA	-0.22	-0.36	-0.03	1.00						1.73

Board Gender Diversity and Social engagement: evidence from the banking industry

Leverage	0.21	0.05	0.15	0.12	1.00					1.52
Coverage	0.04	-0.09	-0.05	-0.01	0.10	1.00				1.04
LoanDep	0.02	-0.11	-0.05	-0.07	0.28	0.07	1.00			1.37
Capex	-0.20	-0.04	-0.1	0.50	0.09	-0.05	-0.03	1.00		1.40
GDPper	0.18	0.44	0.11	-0.34	-0.20	-0.01	-0.14	-0.22	1.00	1.38

***Table 3**: FE regression with SOC pillar score.*

	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Variables	***SOC (t)***							
BoardGD (t-1)	0.13** (3.20)	0.13** (3.02)	0.12** (2.94)	0.12** (2.88)	0.11** (2.61)	0.13** (2.83)	0.12** (2.81)	0.13** (2.84)
GDPper (t-1)	0.00** (2.18)	0.00* (1.94)	0.07** (2.70)	0.00** (2.16)	0.00 (1.25)	0.00 (1.24)	0.00 (1.23)	0.00 (1.26)
B_ind (t-1)		0.07** (2.79)	0.00** (2.00)	0.07** (2.72)	0.07** (2.67)	0.05* (1.71)	0.05* (1.72)	0.05* (1.72)
SustComp (t-1)			2.31** (2.84)	2.22** (2.72)	2.19** (2.71)	1.50* (1.78)	1.51* (1.79)	1.48* (1.75)
logTA (t-1)				1.85 (1.31)	2.55* (1.81)	1.73 (1.17)	1.77 (1.19)	1.73 (1.17)
Leverage (t-1)					-0.13*** (-4.52)	-0.13*** (-4.17)	-0.14*** (-4.16)	-0.14*** (-4.19)
Coverage (t-1)						-0.00 (-1.25)	-0.00 (-1.26)	-0.00 (-1.25)
LoanDep (t-1)							0.00 (0.29)	0.00 (0.30)
Capex (t-1)								1.37 (1.49)
_cons	51.48*** (15.86)	47.07*** (13.05)	46.44*** (12.90)	30.04* (2.31)	35.52** (2.75)	43.65** (3.18)	43.15** (3.12)	43.22** (3.12)
N. of Obs.	1466	1425	1425	1425	1425	1294	1294	1290
R-squared	0.35	0.36	0.37	0.37	0.38	0.38	0.38	0.38

* Asterisks denote statistical significance at the 1% (∗∗∗), 5% (∗∗), and 10% (∗) levels.

5. Conclusions

Gender equality or parity requires a deeper engagement of governments and businesses, with relevant effects on CSR engagement of organizations. The current study focuses on these issues, by advancing existing knowledge on gender balance in banks' top management teams. According to the stakeholder theory and the

conflict resolution hypothesis (Freeman, 1984; Jo and Hrjoto, 2015), this study focused on the link between female banks' board representation and social engagement. The results show that BGD is a significant driver of higher social performance of banks. Given the spread of CSR themes within financial institutions and practices, the current study should encourage banks to increase female representation within corporate boards to improve the engagement toward social issues, and benefit of this greater engagement in terms of loyalty and reputation.

References

1. Alazzani, A., Hassanein, A., Aljanadi, Y.: Impact of gender diversity on social and environmental performance: evidence from Malaysia. Corp. Gov. (2017)
2. Aramburu, I. A.: The effects of corporate social responsibility on customer loyalty: The mediating effect of reputation in cooperative banks versus commercial banks in the Basque country. J. Bus. Ethics **3**, 701-719 (2019)
3. Bear, S., Rahman, N., Post, C.: The impact of board diversity and gender composition on corporate social responsibility and firm reputation. J. Bus. Ethics **97(2)**, 207-221 (2010)
4. Boulouta, I.: Hidden connections: The link between board gender diversity and corporate social performance. J. Bus. Ethics **113(2)**, 185-197 (2013)
5. Brammer, S., Millington, A., Rayton, B.: The contribution of corporate social responsibility to organizational commitment. Int. J. Hum. Resour. Manag. **18(10)**, 1701-1719 (2017)
6. Cosma, S., Leopizzi, R., Pizzi, S., Turco, M.: The stakeholder engagement in the European banks: Regulation versus governance. What changes after the NF directive?. Corp. Soc. Responsib. Environ. Manag. **28(3)**, 1091-1103 (2021)
7. Eagly, A.: Sex differences in social behaviour: A social role interpretation. Hillsdale, NJ: Erlbaum (1987)
8. European Commission: Covid-19: the need for a gender response (2021). Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/689348/EPRS_BRI(2021)689348_EN.pdf
9. Freeman, R.: Strategic management: A stakeholder approach (p. 46). Boston: Pitman (1984)
10. Galletta, S., Mazzù, S., Naciti, V., Vermiglio, C.: Gender diversity and sustainability performance in the banking industry. Corp. Soc. Responsib. Environ. Manag. (2021)
11. Gangi, F., Meles, A., D'Angelo, E., Daniele, L. M.: Sustainable development and corporate governance in the financial system: Are environmentally friendly banks less risky?. Corp. Soc. Responsib. Environ. Manag. **26(3)**, 529-547 (2019)
12. Goss, A., Roberts, G. SJ.: The impact of corporate social responsibility on the cost of bank loans. J. Bank. Financ. **35(7)**, 1794-1810 (2011)
13. Harjoto, M., Laksmana, I., Lee, R.: Board diversity and corporate social responsibility. J. Bus. Ethics **132(4)**, 641-660 (2015)
14. Huse, M., Nielsen, S. T., Hagen, I. M.: Women and employee-elected board members, and their contributions to board control tasks. J. Bus. Ethics **89(4)**, 581-597 (2009)
15. Jain, T., Jamali, D.:. Looking inside the black box: The effect of corporate governance on corporate social responsibility. Corp. Gov. **24(3)**, 253-273 (2016)
16. Lu, J., Herremans, I. M.: Board gender diversity and environmental performance: An industries perspective. Bus. Strategy Environ. **28(7),** 1449-1464 (2019)
17. Naciti, V.: Corporate governance and board of directors: The effect of a board composition on firm sustainability performance. J. Clean. Prod. **237**, 117727 (2019)
18. Nielsen, S., Huse, M.: The contribution of women on boards of directors: Going beyond the surface. Corp. Gov. **18(2)**, 136-148 (2010)

Board Gender Diversity and Social engagement: evidence from the banking industry
19. Renneboog, L., Ter Horst, J., Zhang, C.: Socially responsible investments: Institutional aspects, performance, and investor behavior. J. Bank. Financ. **32(9)**, 1723-1742 (2008)
20. Shakil, M. H., Tasnia, M., Mostafiz, M. I.: Board gender diversity and environmental, social and governance performance of US banks: Moderating role of environmental, social and corporate governance controversies. Int. J. Bank Mark. (2020)
21. Scholtens, B.: Corporate social responsibility in the international banking industry. J. Bus. Ethics **86(2)**, 159-175 (2009)
22. United Nations, 2017. Available at: https://www.un.org/sustainabledevelopment/gender-equality/
23. WTF (2020) The Global Gender Gap Report. https://www3.weforum.org/docs/WEF_GGGR_2020.pdf
24. Wu, M. W., Shen, C. H.: Corporate social responsibility in the banking industry: Motives and financial performance. J. Bank. Financ. **37(9)**, 3529-3547 (2013)

Session of free contributes SCL8 – *Education*
Chair: Matilde Bini

Assessing heterogeneity in students' performance. The case of the Massive Open Online Courses.

Analisi dell'eterogeneità nella performance degli studenti di corsi Massive Open Online.

Cristina Davino and Giuseppe Lamberti

Abstract Exploiting the conceptualization of MOCC performance and one of its main drivers –engagement– this paper aims to evaluate if and how much the role played by engagement in the prediction of student's performance changes according to the students' characteristics –gender, age, and country– or considering different types of courses. To that end we explored the suitability of two tests, the parametric and permutation test, to assess whether the presence of heterogeneity in the sample has different effects on different parts of the conditional distribution of the students' performance. Our results indicate that engagement varies according to the conditional quantiles of the performance and the specific segment defined by the categorical variables for which is estimated.

Abstract *L'obiettivo di questo lavoro è valutare se e quanto il coinvolgimento attivo degli studenti nella frequenza dei MOOC ha effetto sulla performance finale e se tale impatto è differenziato in base alle caratteristiche degli studenti - genere, età e paese di provenienza - o considerando una diversa organizzazione del corso. Il contributo innovativo del lavoro consiste nell'estensione di due test, il test parametrico e il test di permutazione, tipicamente utilizzati per il confronto tra modelli basati su stime ai minimi quadrati nel contesto della regressione quantile. In tale modo si riesce a valutare se la presenza di eterogeneità nel campione ha effetti diversi sulle diverse parti della distribuzione condizionata della performance.*

Key words: MOOC, performance, engagement, heterogeneity, quantile regression, coefficients comparison

Cristina Davino
Department of Economics and Statistics, University of Naples Federico II, e-mail: cristina.davino@unina.it

Giuseppe Lamberti
Department of Economics and Statistics, University of Naples Federico II e-mail: giuseppe.lamberti@unina.it

1 Motivation and reference framework

MOOCs are an increasingly common type of course in education, especially in higher education. The structure and delivery of such courses has a strong impact on the way students attend MOOCs and, inevitably, on their final performance. In the learning analytics framework [11] predicting students' performance in MOOCs can be considered one of the main challenges. The elements affecting students' performance are several, some are specifically related to the learning experience (for example student motivation, learning attitude, engagement), others can be defined as external, related to personal characteristics of the student or to specific features of the course. In this paper we will consider one of the main drivers affecting performance, especially in online courses: engagement. This component is a complex and multidimensional concept linked to the degree to which a student becomes actively and continuously involved in the course activities.

Our contribution extends the study of Carannante, Davino, and Vistocco in 2020 [1], where a structural equation model, in the framework of the composite-based approach ([5]; [7]; [12]), is proposed to measure the main factors affecting students' performance in MOOCs. Exploiting the conceptualization of performance and its main drivers (learning and engagement), this paper focuses on the engagement driver and aims to evaluate if and how much the role played by engagement in the prediction of student's performance changes according to the students' characteristics –gender, age, and country– or considering different types of courses. In essence, we aim to handle a possible heterogeneity in modeling students' performance according to their engagement.

The contribution proposed in this paper also has a methodological value because it aims to find a solution to the well-known problem of model comparison in the case of observed heterogeneity. Several contributions can be found in the literature to assess possible differences between statistical models for subgroups of individuals with different characteristics with respect to stratifying variables namely variables outside the model (a typical case are socio-demographic variables). This paper focuses on the literature for multi-group analysis in Partial Least Squares Path Modeling (among many see [3], [10]), well-known multivariate models where latent variables are estimated through simple and multiple ordinary least squares (OLS) regressions.

This paper aims to extend the traditional approaches in the framework of OLS regression to quantile regression models, i.e. to assess whether the presence of heterogeneity in the sample has different effects on different parts of the conditional distribution of the dependent variable.

Quantile regression ([9]; [4]) models the relationship among explanatory variables and conditional quantiles of a dependent variable without assuming any specific conditional distribution. The main strengths of quantile regression are the possibility to estimate models where either the requirements for mean regression, such as homoscedasticity, are violated or interest lies in the outer regions of the conditional distribution. The empirical analysis on real data proposed in this paper falls into both of these cases because the distribution of the considered dependent vari-

able, i.e. student's performance, is strongly asymmetrical and the aim is obviously to try to improve results especially in case of low performing students.

To analyze heterogeneity in modeling students' performance, we explored the suitability of two different tests: the parametric t-test [8] and the permutation test [2]. These tests follow a multi-group approach [**?**] that consists of separating data into segments according to categorical variables (in our case students' characteristics and courses type), estimating separate models for each category which are then compared to identify significant coefficient differences. The parametric t-test uses a bootstrap re-sampling procedure to evaluate coefficient differences. The coefficients for each segment are calculated in each re-sampling and the bootstrap means and standard error estimates are parametrically tested using a t-test. The permutation test evaluates coefficient differences across segments by applying a permutation procedure. After each permutation, data are reassigned to a group, the coefficients are re-estimated and the differences between them (i.e., permutated differences) are calculated. Permutated differences are finally compared with the original differences and a p-value is calculated as one minus the proportion of the number of times in which the original difference is larger than the permuted one on the total number of permutations.

2 Data, results and discussion

Data presented in this paper refer to 3578 students who attended two courses in Political Science on the FedericaX platform, the EdX MOOCs platform of the "Federica WebLearning" Centre at University of Naples Federico II. Each course was offered in two versions: an instructor–paced version and a self–paced version. The instructor–paced is strictly scheduled, with specific dates for assignments, course materials, exams, and a deadline for learners to complete the course and get a certification. Usually, this modality is integrated into an in–site course delivered in blended mode. For the self–paced version, all of the course materials are available as soon as the course starts, assignments and exams do not have due dates, and therefore a learner can progress through the course at its own speed and pass grade in the course, even without completing all of the course materials. Of the 3578 students, 59.4% were female (40.6% male), just less three quarters (74.2%) were aged less-equal thirty-two years (25.8 % more than thirty-two), 31.1% were Asians and 21.7% were European (52.8% were from other countries). As for the course's version offered on the platform, 73.1% of students followed in the self–paced modality (26.9% instructor–paced).

As described in the previous section, the aim of the study is to analyze whether and how much the students' involvement in the planned MOOC activities impacts on the final performance. Performance was the outcome variable measured as the proportion of the correct answers to a set of questions provided by the teacher. This variable (as shown in Fig. 1), is characterized by a high left-hand asymmetry. This is the reason why the OLS estimation could fail in quantifying correctly the effect of

its driver, engagement. Student's engagement was considered as the predictor and estimated as a latent variable through two sub-dimensions, regularity (how a learner spends her/his time on the platform and how she/he organizes the own learning road map) and no-procrastination (the ability of the learner in organizing the learning processes) [1].

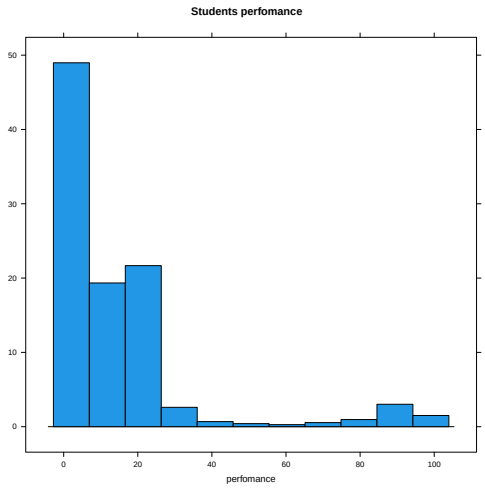


Fig. 1 Student's performance distribution

Comparing the results of the OLS regression and the quantile regression on the entire sample shows different effects of engagement on performance, although always with a positive sign: considering the three conditional quartiles, this effect is increasing (respectively 0.331, 0.431, 0.675) and particularly differentiated in the tails of the distribution compared to the effect on the conditional mean, which is equal to 0.616. To analyze possible differences with regard to student characteristics (gender, age and country) and course type (self-paced, instructor-paced), the same models were estimated on sub-groups of students according to the categories of these external variables.

The effect of engagement on students' performance both in terms of OLS and QR coefficients estimated on the whole sample and for each sub-model is reported in Fig. 2. We graph from left to right the effect of the engagement by students' characteristic country, gender, age, and course type. Each plot reports on the vertical axis the coefficient of the engagement estimated for the quantiles 0.25, 0.5, and 0.7 and for OLS regression. With respect to the QR results we used a black asterisk to indicate the global effect of the engagement; a blue circle and a red triangle to differentiate between the effect of the engagement estimated for each segment defined by the levels of the categorical variables. The full blue circle and the red triangle indicate that coefficients were statically different (p-value < 0.05) according to the permutation and the parametric tests. Further, the black, red, and blue lines

allow appreciating the trend of the coefficients according to the quantiles. Finally, we also report the OLS estimation (global and for each segment) using black plus, the blue square, and the red inverted triangle.

Starting from the top-left panel, we can appreciate that the effect of engagement is significantly higher for students from Asia with respect to the students from Europe when we consider the quantiles 0.25 and 0.5. Interestingly when we observe the quantile 0.75, the relationship reverses in favour of European students but is no longer significant, meaning that the effect of engagement on the top-performing students is similar with respect to the students country. Concerning the effect of gender, the impact of engagement on performance differs by gender in the top 50% of the performance distribution. It is important to highlight the different size of the coefficients in the sub-groups of male and female students at the 0.75 quantile compared to the value obtained from a simple conditional mean estimate. A similar trend can be observed with the variable age. Engagement is significantly higher for older students (more than thirty-two years) when the level of performance is high (quantile 0.75). Finally, concerning the course type the effect of engagement is significantly higher for self–paced students with respect to the instructor–paced but on the worst

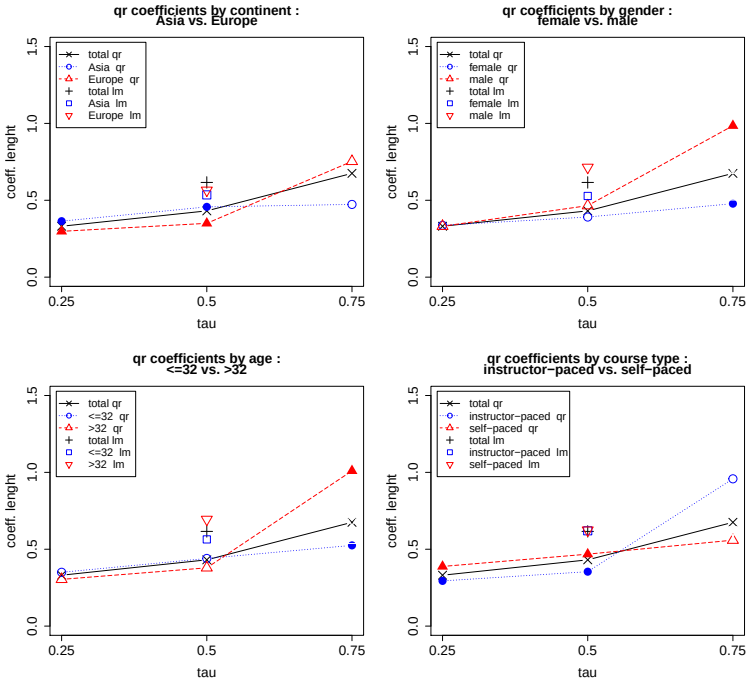


Fig. 2 Engagement quantile regression coefficient by quantile and categorical variables: country, gender, age types of courses

performing students. Again for the quantile 0.75, it could seem that engagement turns to be more important for instructor–paced than for self–paced students; however, the coefficients are not significantly different.

Our results indicate that engagement is an important driver of students' performance. However, its effect is not uniform. Indeed, it varies according to the conditional quantile of the performance and the specific segment defined by stratifying variables not considered in the model. Our findings remark that heterogeneity is a fundamental issue that must be considered to avoid bias results. From the methodological point of view, we show the suitability of the parametric and permutation tests in detecting differences between coefficients in the context of quantile regression. However, future research must be done involving simulation studies to determine better how the tests work and under what conditions. We remind a future study also the improvement of our model by including other predictors of performance as, for example, students learning attitude.

References

1. Carannante, M., Davino, C., Vistocco, D.: Modelling students' performance in MOOCs: a multivariate approach. Stud. High. Educ. **46 (11)**, 2371–2386 (2020)
2. Chin, W. W., Dibbern, J.: An Introduction to a Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services Between Germany and the USA. In: Esposito Vinzi V., Chin W., Henseler J., Wang H. (eds.) Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics, pp. 171-193. Springer, Heidelberg (2010)
3. Chow, G. C.: Tests of equality between sets of coefficients in two linear regressions. Econometrica **28 (3)** 591–605 (1960)
4. Davino, C., Furno, M., Vistocco, D.: Quantile regression: theory and applications. John Wiley & Sons. (2013)
5. Esposito Vinzi, V., Chin, W., Henseler, J., Wang, H. (eds): Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics. Springer, Heidelberg (2010)
6. Hair Jr, J. F., Sarstedt, M., Ringle, C. M., Gudergan, S. P.: Advanced issues in partial least squares structural equation modeling. Sage publications (2017)
7. Hair Jr, J. F., Hult, G. T. M., Ringle, C., Sarstedt, M.: A primer on partial least squares structural equation modeling (PLS-SEM). Sage publications (2016)
8. Keil, M., Tan, B. C., Wei, K. K., Saarinen, T., Tuunainen, V., Wassenaar, A.: A cross-cultural study on escalation of commitment behavior in software projects. MIS quarterly **24 (2)** 299–325 (2000)
9. Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica 33–50 (1978)
10. Sarstedt, M., Henseler, J., Ringle, C. M.: Multi-group analysis in partial least squares (PLS) path modeling: Alternative methods and empirical results. Adv. Int. Mark. **22**, 195–218 (2011)
11. Siemens, G., Long, P.: Penetrating the fog: Analytics in learning and education. EDUCAUSE review **46 (5)**, 30 (2011)
12. Wold, H.: Partial least squares. In S. Kotz e N. Johnson. (eds.) Encyclopedia of Statistical Sciences. John Wiley & Sons (1985)

Assessing undergraduate students' perceptions of distance learning during Covid-19 pandemic

La valutazione della percezione della didattica a distanza degli studenti universitari durante la pandemia Covid-19

Ilaria Primerano, Maria Carmela Catone, Giuseppe Giordano and Maria Prosperina Vitale

Abstract The Covid-19 pandemic has changed the characteristics of university teaching-learning practices marking the transition from face to face to on-line activities by means of digital platforms. This digitization process has produced several effects on students' habits and learning ability. Within this scenario, aim of the contribution is to assess the student perception of the distance learning experience by adopting a network analysis approach. A network of adjectives is obtained from Semantic Differential scales and links are defined as a function of the most frequent polarized items co-occurences. The network core allows to identify the most important adjectives describing the students' e-learning perception.

Abstract *La diffusione del Covid-19 ha modificato i processi di insegnamento-apprendimento universitario segnando il passaggio dalla didattica in presenza a quella on-line, generando diversi effetti sulle abitudini e la abilità di apprendimento degli studenti. All'interno di questo scenario, l'obiettivo del contributo è quello di valutare la percezione degli studenti dell'esperienza di apprendimento a distanza adottando un approccio di network analysis. La rete di aggettivi è ottenuta utilizzando scale di valutazione basate sul Differenziale Semantico e i collegamenti sono definiti in funzione delle co-occorrenze più frequenti e della polarizzazione delle scale. La rete permette di identificare gli aggettivi più importanti che descrivono la percezione dell'e-learning degli studenti.*

Key words: Higher Education, Network Analysis, Semantic Differential, Textual Network

Ilaria Primerano, Giuseppe Giordano, Maria Prosperina Vitale
Department of Political and Social Studies, University of Salerno, Italy e-mail: iprimerano@unisa.it; ggiordano@unisa.it; mvitale@unisa.it

Maria Carmela Catone
Department of Sociology, University of Barcelona, Spain e-mail: mcatone@ub.edu

1 Introduction

The Covid-19 emergency generated a multitude of changes in the University life, such as the implementation of administrative, teaching and research activities in virtual contexts. To ensure the continuity of teaching activities, there was a shift towards forms of distance learning, realized through the use of digital platforms and online resources. The new scenario has lead to a general rethinking of practices, tools and relationships that usually characterize the educational paths.
The use of digital platform to aid university education is not new, taking into account the increasing development over last decades of a wide range of approaches of online education such as e-learning, digital learning, blended learning etc. [1]. At the same time, differently from the "ordinary" circumstances where all the organizational processes are usually scheduled and designed beforehand, the sudden shift of teaching activities in online contexts that occurred during the spread of Covid-19 has taken on forms of "emergency remote teaching" [2], as well as "alternative" modes of educational delivery due to crisis circumstances. Teachers and students in a few days had to modify their traditional modus operandi by changing the times, places and contents of their activities, thus determining significant transformations in terms of practices, social relations, knowledge construction processes, rituals, identity representations, values and beliefs [3].
This configuration has generated a multiplicity of issues that intercept social, pedagogical, disciplinary, technological and economic aspects at micro and macro levels; for example, the owning of digital skills related to the digital divide phenomenona as well as to the perceptions and attitudes of people towards technology. How these unique circumstances have affected the overall level of quality service provided by the University system is a big issue and matter of discussion at scientific and political level. Our proposal move from innovative statistical methodologies for the analysis of data collected during the Covid-19 pandemic to study the effect of distance learning on student experience in higher education. One of the first aspect of interest is to detect the emotional impact of E-learning activities among undergraduate students. At this aim, we use a joint strategy that make use of data collected through a semantic differential scale and apply social network analysis tools. A network of adjectives is obtained from Semantic Differential scales and links are defined as a function of the most frequent polarized items co-occurrences. The network core allows to identify the most important adjectives describing the students' e-learning perception. A pilot study is conducted on a cohort of students to give a major insight of the procedure and evaluate the extent of its applicability.

The contribution is organized as follows. Section 2 describes the methodological procedure, focusing on the strategy adopted to map Semantic Differential data into a network of adjectives. The research design and the main results of a pilot study are presented in Section 3; concluding remarks are in the last section.

2 On the definition of the network of adjectives: The methodology

To analyze students' attitudes towards the crisis scenario defined into the context of university distance learning, as it emerged from the restrictive measures enforced to contain the spread of the virus during the Covid-19 pandemic, we wish to focus on several peculiar aspects: the students' perception on the e-learning experience, their ability on the use of the digital platforms to support learning, and their overall satisfaction level about the e-learning services offered by the university system. This study was one of the first attempts to detect students' perception of university activities at the very beginning of the Covid-19 pandemic. Motivated by these issues, we use a strategy which combines Semantic Differential scales [4] with Social Network Analysis [5]. The Semantic Differential (SD) scale is a well established and reliable psychometric tool useful in assessing emotional attitudes toward a concept of interest. The SD question invites respondents to express their emotional perception on a scale between two polarized options in terms of opposite adjectives. Data collected from a survey containing SD scales are effective in proposing connotative meaning of concepts in forms of most frequent patterns of adjectives polarization. Results are often given at aggregate level of respondents with straightforward charts and tables.

Social Network Analysis (SNA) provides a set of theories and methods for the interpretation and analysis of interactions between entities. While social networks are present in all aspects of our lives (friendships, communication, web, and so on) the network concept represents also a powerful metaphor to identify the shape of connections of a wider range of inter-agent forces in complex systems. In a very general way, a network is a set of elements (nodes) along with a set of links connecting pairs of nodes. Thus, the presence of links corresponds to existing relations among nodes. Abstracting from this general view, we wish to use the network analysis framework, applied to the data obtained from Semantic Differential tasks, in order to extract the main definition of the investigated concepts [6]. This approach consists in the definition of a *Network of Adjectives* by looking at the co-occurrences among set of adjectives and considering also the quantitative scores given by a set of respondents in a Semantic Differential task. The output consists of a particular graph, in which the nodes are the adjectives used for the SD scales and the links depend on the scores assigned to each bipolar scale by the set of respondents. Specifically, two adjectives are linked in the resulting graph if they have been highly scored by a high number of respondents. Specifically, in order to build a semantic network of adjectives it is necessary, firstly, to prepare an ad-hoc coding scheme and, secondly, setting a cutting threshold, to bring out the core of the network, defined starting from the most important links among pairs of adjectives.

After initial pre-processing and data coding, we obtain the adjacency matrix, nodes per nodes, i.e. adjectives per adjectives. This matrix shows all the possible intersections existing between the adjectives considered in the SD scales. Actually, a weighted adjacency matrix is derived, in which the weights correspond to the number of respondents who polarize each pair of adjectives.

Once the data structure has been defined, the use of SNA methods allow to visualize the relationships between adjectives and to analyse them through the main centrality measures [7]. These indices provide a quantitative synthesis with respect to the properties of the whole network and of specific nodes. At the same time, the selection of an optimal cutting threshold defined on the weights of the adjacency matrix allows the central structure of the network (the core) to emerge, as it is made up of the most cohesive group of adjectives it helps to the definition of the stimulus under study.
When the concept to be analyzed is multidimensional, this procedure allows to make comparisons among multiple networks. In such cases the data structure could also be read into the framework of Multilayer Networks [8]. Moreover, if the Semantic Differential adjectives used to measure the respondents' perception of several dimensions of a concept are the same, then a Multiplex Network [9] data structure can be derived.

3 First results from a pilot study

In order to show the procedure described in the previous section, we present an application on real data collected through a pilot study. While the declared aim of the study is to evaluate students' perception of the distance learning activities during the Covid-19 lockdown in the University context, we planned a simple pilot survey based on voluntary response sample of respondents, who were directly involved in distance learning activities.
The questionnaire is divided into five sections. In the first one, we gathered socio-demographic data regarding students' age, gender, education, and family background. The second section included some basic questions about their digital skills and the use of ICT devices. The third section comprised several questions on the organization of distance learning provided by the University. The questions in the fourth section regards the use of the tools available on the e-learning platforms, such as chat, video call, and other apps. The last section is devoted to the definition of the latent concept underpinning students' e-learning experience using the Semantic Differential scales. Great attention has been payed to the choice of the suitable adjectives concerning the concept of distance learning. The nine pairs of adjectives used to define the distance learning concept for all the dimensions under investigation are as follows: useless – useful, undemanding – demanding, unpleasant – pleasant, simple – troublesome, unclear – clear, light – cumbersome, ugly – nice, easy – difficult, permissive – rigorous. The same set of items has been used to evaluate four different dimensions: *e-learning activities, online interactions, use of digital platform*, and *self-study*. [1]

[1] The online questionnaire was administered in the months from May to June 2020 to students who experienced the distance learning by attending the online courses in the Bachelor and Master degree in Sociology delivered for the second term of the a.y. 2019/2020 at the University of Salerno, Italy.

Carrying out the proposed procedure, we have obtained four different networks of adjectives, one for each dimension investigated through the survey. Figure 1 shows the cohesive components of the networks that emerge after optimal cutting of the four adjacency matrices. It shows the subset of adjectives jointly selected by most part of respondents with the highest scores. By looking at these core networks the joint and recurrent presence of some adjectives among the different dimensions clearly emerge, as well as the particular role of an adjective that acts as a bridge in the network. For all four concepts, these networks show a positive connotation, especially for *self-study* and *online communication*. In fact, for these two concepts, the connected component of the network is composed by positive adjectives that allow us to define these two aspects as *simple*, *permissive*, *clear* and *useful*. This last adjective assumes a fundamental role in the networks, as it connects adjectives with a positive meaning with negative ones ("bridge"). This characterization is particularly evident when defining distance *learning activities* and *ICT tools*, which have also been defined as *challenging*, *difficult* and *problematic*.

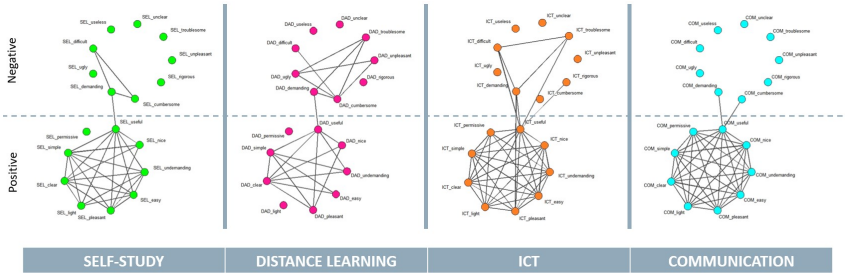


Fig. 1 The four Adjectival Networks emerged from the Semantic Differential to assess students' e-learning perception.

4 Conclusion

Due to the spread of Covid-19, university teaching and learning activities usually carried out in traditional physical spaces have been moved to virtual contexts, using more or less structured forms of distance learning systems. Within this complex configuration, in which disciplinary, technological, economic, social and pedagogical elements converge, it was fundamental to explore the students' point of view as active actors who can constructively contribute to the definition of the learning spaces in which they are involved [10]. In this contribution we presented the first results of a pilot study based on the administration of an online survey to Sociology students of the University of Salerno to explore their attitudes towards their e-learning experience during the first Italian lockdown. The analysis of the Semantic Differential scales using the SNA tools allows to investigate the students' perception. This pro-

cedure allowed us to identify the different connotations characterizing the concept of distance learning. In particular, we found that adjectives with a positive connotations have been chosen by students to define all four concepts related to learning activities (network cores), while in the definition of two concepts, *e-learning* and *ICT*, also adjectives with not exclusively positive connotations emerged.

References

1. Kumar Basak, S., Wotto, M., & Belanger, P.: E-learning, M-learning and D-learning: Conceptual definition and comparative analysis. E-learning and Digital Media **15**(4), 191-216 (2018)
2. Hodges, C.B., Moore, S., Lockee, B.B., Trust, T., & Bond, M.A.: The difference between emergency remote teaching and online learning (2020)
3. Radha, R., Mahalakshmi, K., Kumar, V.S., & Saravanakumar, A.R.: E-Learning during lockdown of Covid-19 pandemic: A global perspective. Int. J. Control. Autom. **13**(4), 1088-1099 (2020)
4. Osgood, C.E., Suci, G.J., & Tannenbaum, P.H.: The measurement of meaning, University of Illinois press (1957)
5. Wasserman, S., & Faust, K. Social network analysis: Methods and applications, Cambridge University Press (1994)
6. Giordano, G., & Primerano, I.: The use of network analysis to handle semantic differential data. Qual Quant **52**(3), 1173-1192 (2018)
7. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215-239 (1978)
8. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., & Porter, M.A.: Multilayer networks. J. Complex Netw., **2**, 203–271 (2014)
9. Dickison, M.E., Magnani, M., & Rossi, L.: Multilayer social networks. Cambridge University Press (2016)
10. Ghislandi, P., & Raffaghelli, J.: La voce degli studenti per la qualità dell'eLearning nella formazione universitaria: un approccio partecipativo. Student Voice. Prospettive internazionali e pratiche emergenti in Italia, 273-286 (2013).

PhD satisfaction analysis in Italian University via Classification tree, Bagging and Random Forest.

Analisi della soddisfazione dei Dottori di ricerca nella Università Italiane attraverso Alberi di classificazione, Bagging e Random Forest

Anna Crisci and Antonio Lucadamo and Pietro Amenta

Abstract In this work we consider the classification methods as trees, bagging and random forest in order to analyze the determinants of PhD student satisfaction in Italian universities. Understanding how doctoral students see their experience can help to improve the quality of programs and to obtain better results, for example in terms of time to degree completion, or the formation of better researchers.
Decision tree results show that the most important variables are those linked to supporting the career at the end of the PhD program.
Abstract: *In questo lavoro consideriamo i metodi di classificazione come alberi decisionali, bagging and random forest al fine di analizzare le determinanti della soddisfazione dei dottorandi nelle università italiane. Capire come i dottorandi vedono la loro esperienza può aiutare a migliorare la qualità dei programmi e ad ottenere risultati migliori, ad esempio in termini di tempo per il completamento del corso o la formazione di ricercatori migliori.*
I risultati dell'albero decisionale mostrano che le variabili più importanti sono quelle legate al sostegno alla carriera al termine del dottorato.

[1]Anna Crisci, department of Economics , Management and Institutions, University of Naples Federico II, email: anna.crisci@unina.it;

Antonio Lucadamo, Department of Law, Economics, Management and Quantitative Methods, University of Sannio, email: antonio.lucadamo@unisannio.it;

Pietro Amenta, Department of Law, Economics, Management and Quantitative Methods, University of Sannio, email: amenta@unisannio.it

Key words: Classification trees, Bagging, Random Forest, PhD-doctoral student satisfaction

1 Introduction

In this work we consider the classification methods as trees, bagging and random forest in order to investigate on the determinants of PhD student satisfaction in Italian universities. Student satisfaction and the quality of education are of engaging importance to students, academic staff, policymakers and researcher. Usually, studies among PhD students have tended to focus on identifying particular issues, such as supervision, research training and employability, and have been assumed largely for internal institutional use and for quality improvement.
In our study, we consider the data collected in the third national "statistical survey of the employability of the doctoral graduates of 2012 and 2014" conducted by Istat (Istat, 2019).
The response variable is the level of satisfaction with the overall PhD experience.
The Key independent variables are:

- Quality of the didactic activities received (Quality_teach) such as: depth and currency of course contents, rigor/adequacy of teaching methods;
- Quantity of training activities received (Quantity_train): seminars, workshops, conferences, winter and summer schools;
- Competence of the academic teaching staff (Competence_Teach). In particular, their expertise can be measured by their teaching abilities, scientific output and academic curriculum;
- Research facilities (Research_facilities): availability and accessibility of individual working spaces, learning resources, laboratories and laboratory equipment;
- Research training (Research_Train): this question relates to the satisfaction with the guidance received during the various stages of the research process which equips the candidate with the necessary research skills for becoming an independent researcher;
- Collaboration with teaching and research staff (Collaboration), that is the satisfaction with the interaction with departmental staff.
- Encouragement (Encouragement) to submit works for publication: this is an indicator of the extent to which the doctoral candidate has been recognised as having reached a certain scientific maturity.

For each of these questions, the participants are asked to give a rating ranging from 0 to 8, where 0 indicates "not at all satisfied" and 8 indicates "completely satisfied". Furthermore, variables as gender, citizen and regularity in completing the course are considered. Finally, in order to choice the most suitable model on the basis of the misclassification rate, we compare the classification methods with models for

analysing data with ordinal responses. Ordinal models considered are: Proportional odds Model, Continuation ratio model, Adjacent category model.

2 Methodology

In this section we briefly describe the models and methods we apply in our analysis. We first introduce the classical logistic model for ordinal variables and then classification methods as trees, bagging and random forest.

2.1 Models for ordinal response

Proportional odds model, Continuation ratio model and Adjacent category model are the most used models in Multinomial Logit when the response variable is ordinal (Agresti, 2007). They differ in the formulation of the logit. In fact in the Proportional odds, we consider the logit between the modalities higher than a value j, and the previous ones. The cumulative probabilities are related to a linear predictor $\boldsymbol{x'\beta}$, through the logit function:

$$\pi_j(x) = \frac{P(Y>j|x)}{P(Y\leq j|x)} = \alpha_j - \boldsymbol{x_i'\beta} \;\; for\; j = 1,2,\dots J-1$$

The parameters α_j, called thresholds or cut points, are of increasing order $\alpha_1 \leq \alpha_2 \dots \leq \alpha_{J-1}$ and $\boldsymbol{\beta}$ is a vector of logit coefficients

In the continuation ratio the logit is between a generic category j and all the modalities with value lower than *j*:

$$\pi_j(x) = \frac{P(Y=j|x)}{P(Y\leq j|x)} = \alpha_j - \boldsymbol{x_i'\beta}\; for\; j = 1,2,\dots J-1$$

In the adjacent category the comparison is between consecutive modalities:

$$\pi_j(x) = \frac{P(Y=j|x)}{P(Y=j-1|x)} = \alpha_j - \boldsymbol{x_i'\beta_j}\; for\; j = 1,2,\dots J-1$$

All models are fitted through the procedure of maximum likelihood estimation (Agresti, 2015).

2.2 Classification tree

Classification tree is used to predict a qualitative response using qualitative or quantitative predictors. The tree is built according to two steps: first of all the predictor space is divided into distinct regions; then, for every observation that falls into each region, the prediction is made, corresponding to the modality of the response with the higher proportion in that region (Hastie et al., 2001). The problem is that it is computationally infeasible to consider every possible partition of the space and for this reason a top-down approach is used. It begins at the top of the tree, where all observations are in the same region, and then the predictor space is splitted successively. At each step, the split is indicated with two branches, selecting the predictor and the cutpoint s such that, splitting the predictor space into the regions $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$, leads to the greatest reduction in the classification rate. The process continues until a stopping criterion is reached. Once the regions have been created, the response is predicted considering the most commonly occurring class in that region.

2.3 Bagging

The regression and classification trees suffer from high variance. In fact, splitting the training data into two parts at random, and fitting a decision tree, the results could be quite different. In contrast, a procedure with low variance will yield similar results if applied repeatedly to distinct data sets. Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method; it is particularly useful and frequently used in the context of decision trees. To reduce variance and so increase the accuracy of prediction, a way is to take many B training sets, build a separate prediction model using each training set and then average the resulting predictions (James et al., 2013):

$$f_{bootstrap}(x) = \frac{1}{B}\sum_{b=1}^{B} f^b(x)$$

where, $f^b(x)$ is the predictive result of a specific training set.

In the case of classification problem the overall prediction is the most commonly occurring class among B separate training sets.

2.4 Random Forests

In random forests, as in bagging, decision trees on bootstrapped training samples are built, but in this case a random sample of m predictors is chosen, as split candidates from the full set of predictors. This is done because if there is one very strong predictor in the data set, along with a number of other moderately strong predictors, then in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Consequently, all of the bagged trees will look quite similar to each other and this means that bagging will not lead to a substantial reduction in variance over a single tree in this setting. Random forests overcome this problem by forcing each split to consider only a subset of the predictors.

The predictor based on the Random Forest model is:

$$f_{RF}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_B\,(x, \vartheta)$$

where b_i indicates the tree of the Random Forest model and the term $T_B(x, \vartheta)$ is the prediction of the same tree.

3 Results

In this section we compare different methods on the basis of the misclassification rate. The classification tree leads to a very high rate of misclassification and this result can be improved by considering the various techniques as bagging, random forest, that can increase the accuracy of the model.

Table 1: *Misclassification rate for the different methods.*

Model/Method	***Misclassification rate***
Proportional odds model	0.5138
Adjacent category model	0.5013
Continuation ratio model	0.4966
Classification tree	0.5602
Bagging	0.0882
Random Forest	0.0633

In table 1 the comparison between ordinal logistic regression model, Classification tree, Bagging and Random Forest, shows a clearly lower rate of misclassification for the random forest and bagging model. Table 2 shows instead, the most important variables on the basis of the mean decrease in node impurity. The most important variables are therefore: Research training, Research facilities, Quality of the teaching, Encouragement to submit works for publication.

Table 2: *Variable Importance and rank for Classification Tree, Bagging and Random forests*

Model/Method	***Classification tree***	***Bagging***	***Random forest***
Quality teach	1313.36 (1)	1061.60 (3)	879.97 (4)
Quantity train	701.53 (5)	850.31 (7)	801.58 (7)
Competence teach	471.36 (7)	973.29 (4)	826.84 (6)
Research facilities	536.44 (6)	1153.90 (2)	917.19 (2)
Research train	1229.72 (2)	1236.16 (1)	936.24 (1)
Encouragement	741.99 (4)	962.11 (5)	913.28 (3)
Collaboration	904.28 (3)	910.05 (6)	899.45 (5)
On Time*		306.63 (9)	229.42 (9)
Gender		451.45 (8)	300.93 (8)
Citizen**		95.99 (10)	84.38 (10)

*On time: Has concluded the programme on time, Yes – No
**Citizen: Italian or foreign

References

1. Agresti, A.: An introduction to categorical data analysis. Wiley (2007).
2. Agresti, A.: Foundations of Linear and Generalized Linear Models. Wiley (2015).
3. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Data Mining, Inference and Prediction. Springer (2001)
4. Istat (2019).: Indagine sull'inserimento professionale dei dottori di ricerca, Anno 2018. Istituto Nazionale di Statistica.
5. James, G, Witten, D., Hastie, T., Tibshirani, R.: An introduction to Statistical Learning with application in R. Springer, Heidelberg (2013)

Comparative Analysis of Student Learning: Technical, Methodological and Result Assessing of PISA-OECD and INVALSI-Italian Systems

Analisi Comparata dell'Apprendimento degli Studenti: Valutazione di Tecnica, Metodologia e Risultati dei Sistemi PISA dell'OCSE e INVALSI Italiano

S. Cervellera, C. Cusatelli and M. Giacalone

Abstract PISA is the most extensive international survey promoted by the OECD in the field of education, which measures the skills of fifteen-year-old students from more than 80 participating countries every three years. INVALSI are written tests carried out every year by all Italian students in some key moments of the school cycle, to evaluate the levels of some fundamental skills in Italian, Mathematics and English. Our comparison is made up to 2018, the last year of the PISA-OECD survey, even if INVALSI was carried out for the last edition in 2022. Our analysis focuses attention on the common part of the reference populations, which are the 15-year-old students of the 2nd class of secondary schools of II degree, where both sources give a similar picture of the students.

Abstract *PISA è la più ampia indagine internazionale promossa dall'OCSE nel campo dell'istruzione, che misura ogni tre anni le competenze di studenti quindicenni provenienti da più di 80 paesi partecipanti. Le INVALSI sono prove scritte svolte ogni anno da tutti gli studenti italiani in alcuni momenti chiave del ciclo scolastico, per valutare i livelli di alcune competenze fondamentali in Italiano, Matematica e Inglese. Il nostro confronto è effettuato fino al 2018, ultimo anno dell'indagine PISA-OCSE, sebbene l'ultima edizione dell'INVALSI sia del 2022. La nostra analisi focalizza l'attenzione sulla parte comune delle popolazioni di riferimento, che sono gli studenti quindicenni della 2ª classe delle scuole secondarie di II grado, dove entrambe le fonti forniscono un quadro simile degli studenti.*

Key words: Student Learning, PISA-OECD, INVALSI

[1] S. Cervellera, Municipality of Taranto; email: s.cervellera@comune.taranto.it
C. Cusatelli, University of Bari "Aldo Moro"; email: carlo.cusatelli@uniba.it
M. Giacalone, University of Naples "Federico II"; email: massimiliano.giacalone@unina.it

1 Introduction

With this paper we have the goal of comparing the methodology and results of the two main surveys on the Italian students' learning, represented by the OECD PISA (Organisation for Economic Cooperation and Development's Programme for International Student Assessment) survey and the Italian one, called INVALSI (National Institute for the Assessment of the Educational System of Learning and Training). The comparison is made up to 2018, the last year of the PISA-OECD survey, even if INVALSI was carried out for the last edition in 2022 (ex 2021 postponed due to health emergency).

2 The programme for international student assessment

PISA is an international survey promoted by the OECD which measures the skills of fifteen-year-old students from the participating countries every three years, the most extensive international survey in the field of education, attended by students from more than 80 different countries [2-5]. PISA-OECD tests are structured for the purpose of detecting some skills of students when they are about to finish compulsory schooling, with the aim of monitoring these students when they should be ready to face adult life, focusing attention on the mastery of curricular contents, which are evaluated by individual school systems and which are difficult to compare with each other. The focus is on the ability to face and solve the problems of everyday life and the ability to be able to continue learning in the future. PISA results therefore allow schools, education systems and governments to identify aspects to improve in their educational programs, to train more competent citizens. They also allow you to compare student performance and learning contexts in different countries.

To determine students' problem solving and lifelong learning skills, PISA measures the skills of 15-year-olds, and focuses on evaluating students' performance in Reading, Mathematics and Science because the skills underlying the study of these subjects are fundamental to face adult life. PISA tests are mainly made up of multiple choice questions, but also include open-ended questions, which can make up up to one third of the test, as well as a basic questionnaire, providing information about themselves and their attitude towards learning. School administrators also receive a questionnaire on their schools that integrates the information provided by the students, while some countries may, independently, also decide to administer other optional PISA questionnaires: a questionnaire on familiarity with the computer, one on the educational career and one on the background cultural heritage of the parents. Each edition then includes an optional survey on a further topic. The 2015 edition, for example, deepened the collaborative problem solving ability.

PISA tests are held every three years, starting from 2000 and at each investigation one of the three areas that are subject to measurement is deepened. In the last edition of 2018, for example, the main domain was the reading literacy

which refers to the understanding, use and reflection on written texts in order to achieve one's goals, the ability to develop one's knowledge and potential and to play an active role in society.

The results of each survey are usually published the year following that of administration. The last edition of the PISA Survey took place in 2018 and its results are available from 3 December 2019 and the 2022 edition is already being prepared, postponed for a year due to the health emergency, in which the 36 OECD member countries and about 50 other non-member countries will participate. The focus of the three-year period will be on Mathematics and the students will also be tested on a new discipline: creative thinking. For the 2025 edition it has already been established that the focus will be on Sciences and that the skills of learning in a digital world will be tested.

3 The national institute for the assessment of the educational system of learning and training

INVALSI tests are written tests carried out every year by all Italian students of the classes provided for by the legislation. Their purpose is to evaluate, in some key moments of the school cycle, the levels of learning of some fundamental skills in Italian, Mathematics and English that the legislation provides are possessed by all students. Based on the elaboration of the test results, indications are obtained for evaluation at class, institute, regional and national level [1]. The first proposals for a National System of Evaluation of the school system date back to the early nineties of the twentieth century, in conjunction with the discussion of the reform that will introduce the autonomy of educational institutions.

As at the international level, in Italy evaluation is considered an important element of the evolution of the school system towards greater effectiveness and adaptation to the needs of a rapidly changing social, cultural and economic context. This last requirement was also the basis of the National Indications and Guidelines, the documents that have replaced the old didactic programs, on the basis of which INVALSI elaborates the theoretical references and the operational aspects of the tests, starting from the skills to be evaluated.

After two school years of experimentation, the first INVALSI National Tests of Mathematics and Italian were held in the 2005-06 school year and since then, the tests have undergone changes partly dictated by the legislation, which has changed several times the choice of the moments of the school cycles in which to carry out the evaluation, partly introduced following the continuous improvement of the quality of the questions, the methods of administration, and the methods of statistical processing of the results, made possible by the research activity. INVALSI tests are regulated by Legislative Decree no. 62 of 2017, following which important innovations were introduced, and from 2018, in the fifth primary and in the third secondary of the first degree, to the Mathematics and Italian tests an English one has also been added which includes a reading and a listening test. From the same year,

while in the second and fifth primary the tests are still carried out on paper file, in all other grades they are carried out using computer based testing and the results are transferred directly to INVALSI.

Measuring in a standardized way some fundamental skills in Italian, Mathematics and English, the results are comparable between different schools or geographical areas and, thanks to statistical techniques of "anchoring", from one year to the next. The competences examined are some of those that the legislation provides to be taught and learned in Italian schools starting from the National Indications and the Guidelines of the various classes. In fact, INVALSI elaborates the Reference Frameworks for assessment, documents that also take into account similar international documents and teaching practice, and on the basis of which the authors of the tests work, because not only knowledge but also skills of students are examined: INVALSI are not memory tests but measure the ability to think about real-life issues or problems, to use the knowledge learned, to connect them with each other and to apply them to new problems.

With the results obtained from the tests it is possible to identify any strengths and difficult situations, but also to discover any inequalities from school to school or from territory to territory, in addition to the positive or negative dynamics of the results of the time. Examining the results can help to better understand some problems such as early school leaving, gender differences, the inclusion of foreign pupils, or the effect of schools on the preparation of students along the school cycle. The results of the tests are in fact one of the elements available for the self-assessment activities of educational institutions. The results are not elements for the evaluation of individual students, which remains the exclusive prerogative of the teacher: they do not serve to evaluate the work of teachers, nor does any rewarding or penalizing mechanism for institutions or teachers depend on the results of the tests.

4 PISA and INVALSI comparison

Our analysis focuses attention on the common part of the reference populations, which are the 15-year-old students of the 2nd class of secondary schools of II degree, where both sources give a similar picture of the performance differentials between students. The picture relating to the overall variability is also similar, although critical issues from this point of view are found in the INVALSI, due to the presence of a significant presence of cheating phenomena and an imperfect correction of the same, which generates a further source of variability in the results. The consistency between the two sources extends to the results of the individual students and especially of individual schools. These evidences constitute a good viaticum for a joint use – and redesign – of the two surveys. In particular, they suggest the desirability of an ex ante design of mechanisms for anchoring INVALSI results in the metric provided by PISA, with the advantage of allowing comparisons in time and space, compared to other OECD countries.

PISA-OECD is a three-year sample survey referring to 15-year-old students, regardless of the school level pursued, while the national survey on learning is conducted through the INVALSI national sample survey, with annual frequency. Both are aimed at the universe of students of a certain school grade. This also results in differences in the structure of the two surveys: PISA-OECD based on a wider set of questions, in rotation proposed to the different individual participants, with higher levels of control with a greater number of open questions, which is carried out with specially trained, trained and paid staff. INVALSI is based on a narrower set of questions equally asked to all individual participants, with fewer open questions and external control systems only in subsets of classes.

The aims of the two statistical surveys are different: PISA-OECD aims to compare the Italian system with the rest of the world (international/comparative character), while INVALSI as feedback to the individual school on its performance and geographical ranking of Italian school efficiency. The main analysis is to verify if the overall picture that emerges from the two surveys is consistent, both in methodological terms and in terms of results. It would be profitable to have a finer monitoring of the internal events of the country and to take advantage of the greater coverage of the INVALSI annual surveys, and correlate them to an international evaluation system with the rest of the world, with the PISA methodology, which would make the comparison also assessable with other international surveys that take place in other school grades.

5 Some comparison results

The areas of specialization in common for both surveys are Reading and Mathematics in PISA, corresponding in INVALSI to Italian and Mathematics. The comparison will be carried out both in terms of the general pattern of the results and its main regularities such as the territorial gap, the structural differences and the characteristics of the students and considering the results at the level of the individual schools. The differences in the format of the two surveys will be evaluated, differences whose possible impacts are not easily identifiable and eliminable, but which must also be kept in mind when interpreting the results of the comparison.

Through the use of standardized variables for years and type of survey, it is possible to better compare and evaluate the result divergences in the two survey as well as differences in the same for the different territorial divisions:

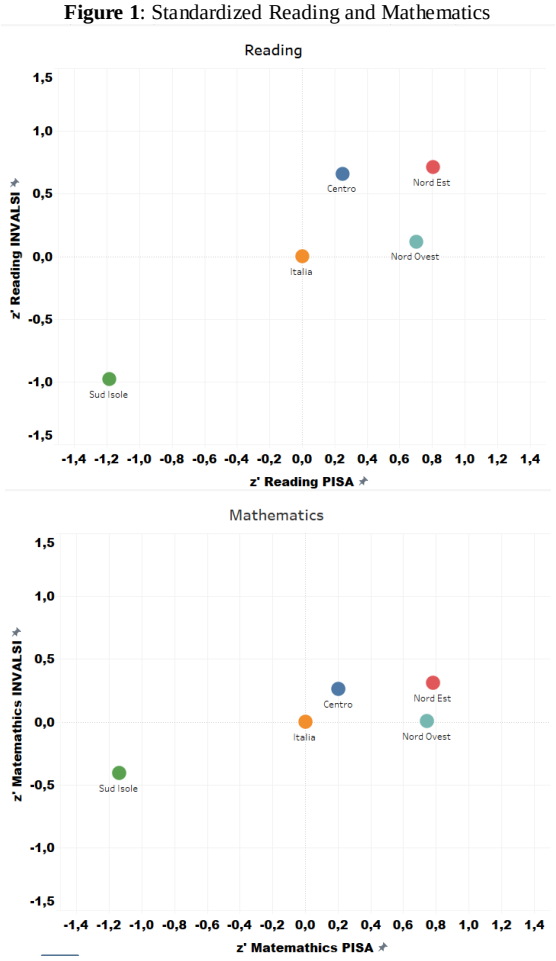
$$z_{t,y,d,s} = \frac{x_{t,y,d,s} - \mu_{t,y,s}}{\sigma_{t,y,s}}$$

where t = PISA or INVALSI test, y = year, d = Italian territorial division, s = school subjects.

Fig. 1 shows an increase in the divergence between PISA and INVALSI general surveys, which intensifies in the territorial level of distribution, in particular in the

North-South divide, representing a greater discordance in Reading and much less in Mathematics, also at the level of territorial divisions.

Figure 1: Standardized Reading and Mathematics



References

1. INVALSI, OCSE: PISA 2018 - I risultati degli studenti italiani in Lettura, Matematica e Scienze, Rapporto nazionale, Area indagini internazionali INVALSI (2018)
2. OECD: PISA 2018 - Results (Volume I): What Students Know and Can Do, PISA, OECD Publishing, Paris (2019)
3. OECD: PISA 2018 - Results (Volume II): Where All Students Can Succeed, PISA, OECD Publishing, Paris (2019)
4. OECD: PISA 2018 - Results (Volume III): What School Life Means for Students' Lives, PISA, OECD Publishing, Paris (2019)
5. OECD: PISA 2018 - Results (Volume IV): Are Students Smart about Money?, PISA, OECD Publishing, Paris (2020)

Session of solicited contributes SS7 – *SEM with PLS: Theory and Applications*
Organizer and Chair: Enrico Ciavolino

PLS-SEM basics and its potential applications: A Quick Journey

Nozioni di base su PLS-SEM e sue potenziali applicazioni: un viaggio veloce

Siqi Wang, Jun-Hwa Cheah, José L. Roldán[1]

Abstract Partial least squares based structural equation modeling (PLS-SEM) has gained popularity as one of the preferred multivariate data analysis methods for a variety of analysis scenarios. This study hopes to disseminate more current PLS-SEM techniques or practices to reach a wider audience. These techniques include conditional mediation model, prediction using PLSpredict or CVPAT, confirmatory purpose with model fit, and model selection criteria. This study will benefit novice researchers or encourage researchers to try these interesting techniques to assist in their research.

Abstract *La modellazione delle equazioni strutturali basata sui minimi quadrati parziali (PLS-SEM) ha guadagnato popolarità come uno dei metodi di analisi dei dati multivariati preferiti per una varietà di scenari di analisi. Questo studio spera di diffondere tecniche o pratiche PLS-SEM più attuali per raggiungere un pubblico più ampio. Queste tecniche includono il modello di mediazione condizionale, la previsione mediante PLSpredict o CVPAT, lo scopo di conferma con l'adattamento del modello e i criteri di selezione del modello. Questo studio andrà a beneficio dei ricercatori alle prime armi o incoraggerà i ricercatori a provare queste tecniche interessanti per aiutare nella loro ricerca*

Key words: Conditional mediation model, PLSpredict, CVPAT, confirmatory purpose with model fit, model selection criteria

[1] Siqi Wang, School of Business and Economics, Universiti Putra Malaysia, Selangor, Malaysia; email: rubywangsiqi@gmail.com

Jun-Hwa Cheah, School of Business and Economics, Universiti Putra Malaysia, Selangor, Malaysia; email: jackycheahjh@gmail.com

José L. Roldán, Department of Business Administration and Marketing, Universidad de Sevilla, Av. Ramón y Cajal, 1, 41018 Seville, Spain; email: jlroldan@us.es

1 Introduction

Structural Equation Modeling (SEM) is a multivariate technique that combines both principal component features and regression analysis (Hair et al., 2017a; Sarstedt et al., 2017). Researchers have begun to recognize their ability to model latent variables, take into account various forms of measurement error, and test grounded theories in a structured manner (Pakpahan et al., 2017). There are two methods of SEM: covariance based structural equation modeling (CB-SEM; also known as component based SEM) and partial least squares based structural equation modeling (PLS- SEM; also known as composite-based SEM) (Hair et al., 2021). CB-SEM is a suitable method to use when the study is aimed at theoretical testing and confirmation. CB-SEM follows a common factor model that assumes that the observed scores from the metrics are a function of the construct itself and the measurement error. To estimate model parameters such as indicator loadings and path coefficients, the method uses only the common variance (i.e., the variance shared by the construct indicators) (Hair et al., 2021). The focus of CB-SEM has been primarily on small conceptual models, which has hindered the development and validation of large, complex models (Chin, Peterson & Brown, 2008). In contrast, for prediction and theory development, the appropriate approach is PLS-SEM. PLS-SEM follows a composite model logic that uses total variance and represents the construct as a linear combination of its indicators (Sarstedt et al., 2016). The goal is to predict and explain a key target structure and/or identify its associated antecedent structure (Chin et al., 2020). Because of these uniqueness, the technique has been growing among multi-disciplinary, such as in accounting, marketing, information systems, and psychology. Given the recent research and methodological developments in the field of PLS-SEM, there is a need for continuous dissemination of the latest technique to other disciplines of research. Therefore, this short article provides some interesting basic and potential applications to scholars when examine their proposed model using PLS-SEM.

2 Available Technique in PLS-SEM

In these section, we explained four potential applications of PLS-SEM techniques that could enhance or capture better findings in their research modeling. These are:

2.1 Conditional Mediation Model

Conditional mediation (CoMe) analysis denotes the statistical assessment carried out when analyzing and estimating a CoMe model (Cheah et al., 2021). CoMe analysis combines mediation and moderation analyses to examine and test hypotheses about

how mediated relationships vary because of context, boundaries, or individual differences. It occurs when a moderator interacts with one or more of the paths of the mediated effect, such that the value of the mediated effect changes depending on the value of the moderator (Hayes, 2017, 2018). This type of modeling is well suited to investigate how relationships between cause and outcomes vary depending on the characteristics of their contexts (Bachl, 2017).

The use of PLS-SEM for estimating a CoMe model offers the following advantages: (1) it overcomes the limitations of traditional sequential approaches by enabling researchers to analyze complex interrelationships between latent variables simultaneously, (2) while accounting for the measurement error inherent in the multi-item measurements (Edwards & Lambert, 2007; Hayes & Scharkow, 2013; Muller et al., 2005). Such modeling provides deeper insights into the intricacies of processes or under which conditions they occur.

Overall, CoMe analysis often opens up new avenues for analyzing new research questions. To recap the utility of conducting CoMe analysis in PLS-SEM, the study by Cheah et al. (2021) provide three concluding recommendations that are concerning model specification, causal inferences, and issues regarding sample size.

2.2 Prediction using PLSpredict or CVPAT

Shmueli et al. (2016) developed the PLSpredict procedure for generating holdout sample-based point predictions in PLS path models on an item or construct level. Liengaard et al. (2020) established the cross-validated predictive ability test (CVPAT) method, which is non-parametric. The purpose of this new method is to conduct a pairwise comparison between two theoretically derived models regards their ability to predict the indicators for all the dependent latent variables (regardless whether reflective or formative) simultaneously.

Both the PLSpredict and CVPAT techniques allow researchers to address the long-standing calls for a stronger focus on predictive model assessment, most notably a model's out-of-sample predictive power (Liengaard et al., 2020; Shmueli et al., 2016, 2019). By having low prediction errors (e.g., using the root mean square error (RMSE) and the mean absolute error (MAE) statistic) and a high value of Q2_predict, researchers can identify a parsimonious model that is more likely to predict and be generalizable to other samples. Similarly, having appropriate pairwise comparison results for CVPAT (see Chin et al., 2020) with its overall inferential test enables researchers to statistically compare the predictive strengths of models to judge whether model choice is reliable, and not affected by the chance of sampling error. Importantly, these criteria enable practitioners to make such decisions with less error by reducing generalization error so that policy decisions will be more likely to work in other settings.

2.3 Confirmatory Purpose with Model Fit

Fit measures, such as the standardized root mean square residual (SRMR) and bootstrap-based tests of the model fit (the unweighted least squares (dULS) and the geodesic discrepancy (dG)), play an important role in guiding researchers to assess whether the data follow a common factor model or composite model, when determining the characteristic of the construct to be measured (Dijkstra & Henseler, 2015; Sarstedt et al., 2016). If the specific measurement does not meet the required level of goodness of fit, this denotes that the data may exhibit the characteristics of a composite model (Sarstedt et al., 2016). In addition, when research goal is to achieve confirmatory in a particular study, the confirmatory purpose using fit indexes can be applicable both with PLS and PLSc, depending on if we have a component-based model (confirmatory composite analysis) or a factor-based model (confirmatory factor analysis) (Henseler & Schuberth, 2020).

2.4 Model Selection Criteria

Hair et al. (2017b) highlighted that the latest PLS-SEM toolbox included a broad range of evaluation criteria for assessing the adequacy of a model. There were researchers who explored whether the use of in-sample measures, such as the model selection criteria (i.e., Bayesian information criterion (BIC), Geweke–Meese criterion (GM)) could be a potential substitute for out-of-sample criteria that require a holdout sample (Sharma et al., 2019, 2021). That same year, Danks et al. (2020) extended the use of the model selection criteria by looking into the BIC weights (BICw) and GM weights (GMw). In addition, there are several variations of the original BIC criteria that have also been proposed in recent decades, including the Hannan–Quinn Criterion (HQ) and the corrected Hannan–Quinn Criterion (HQC) (see Sharma et al., 2019, 2021). These new criteria were intended to assist scholars in overcoming selection uncertainty when selecting an appropriate model over others alternative model based on the model selection criteria. Importantly, Sharma et al. (2019, 2021) highlighted that the model selection criteria (particularly BIC and GM) are known as in-sample criteria that could be a substitute for out-of-sample criteria that require a holdout sample. Such a substitution is advantageous, especially when the researcher does not have the luxury of a holdout sample (using an insufficient sample for the holdout sample causes considerable loss of statistical and predictive power), and the goal is to select correctly specified models with low prediction error. Subsequently, these model selection criteria help compare different model configurations that could result from different theories or research contexts.

Chin et al. (2020) encourages the use of BICw and GMw, which can be interpreted as conditional probabilities for models (e.g., Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004), thereby offering stronger evidence for or

against each model in the set (Danks et al., 2020). Thus, the use of both the BICw and GMw criteria facilitate researchers in overcoming the false sense of confidence that occurs when selecting between models with similar BIC and GM values.

3 Conclusion

This study shares a quick guide to the basics of PLS-SEM and its potential applications. The "family members" of PLS-SEM are introduced, providing a concise explanation and guidelines. The goal is to understand what these techniques are intended to achieve and to lay the foundation for future applications. Therefore, we encourage researchers to match these techniques to their research questions and designs, and further explore the applicability of each technique in different research areas, thereby helping the academic field move forward.

References

1. Bachl, M.: Conditional Process Modeling (Mediation Analysis, Moderated Mediation Analysis, Moderation Analysis, and Mediated Moderation Analysis). The International Encyclopedia of Communication Research Methods. 1--26 (2017)
2. Burnham, K. and Anderson, D.: Model Selection and Multi-Model Inference, Springer, Heidelberg (2002)
3. Cheah, J. H., Nitzl, C., Roldán, J. L., Cepeda-Carrion, G., & Gudergan, S. P.: The Data Base for Advances in Information Systems. A Primer on the Conditional Mediation Analysis in PLS-SEM. The Data Base for Advances in Information Systems, In Press (2021)
4. Chin, W. W., Peterson, R. A., & Brown, S. P.: Structural equation modeling in marketing: Some practical reminders. Journal of marketing theory and practice. **16**(4), 287--298 (2008)
5. Chin, W., Cheah, J. H., Liu, Y., Ting, H., Lim, X. J., & Cham, T. H.: Demystifying the role of causal-predictive modeling using partial least squares structural equation modeling in information systems research. Industrial Management & Data Systems (2020)
6. Danks, N. P., Sharma, P. N., & Sarstedt, M.: Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLS-SEM). Journal of Business Research. **113**, 13--24 (2020)
7. Dijkstra, T. K., & Henseler, J.: Consistent partial least squares path modeling. MIS quarterly. **39**(2), 297--316 (2015)
8. Edwards, J. R., & Lambert, L. S.: Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis. Psychological methods. **12**(1), 1--22 (2007)
9. Hair, J. F., Astrachan, C. B., Moisescu, O. I., Radomir, L., Sarstedt, M., Vaithilingam, S., & Ringle, C. M.: Executing and interpreting applications of PLS-SEM: Updates for family business researchers. Journal of Family Business Strategy. **12**(3), 100392 (2021)
10. Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M.: A primer on partial least squares structural equation modeling (PLS-SEM). SAGE, Thousand Oaks, CA (2017b)
11. Hair, J. F., Matthews, L. M., Matthews, R. L., & Sarstedt, M.: PLS-SEM or CB-SEM: updated guidelines on which method to use. International Journal of Multivariate Data Analysis. **1**(2), 107--123 (2017a)
12. Hayes, A. F.: Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. 2nd ed. Guilford Publications (2017)
13. Hayes, A. F.: Partial, conditional, and moderated moderated mediation: Quantification, inference, and interpretation. Communication monographs. **85**(1), 4--40 (2018)

14. Hayes, A. F., & Scharkow, M.: The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter?. Psychological science. **24**(10), 1918--1927 (2013)
15. Henseler, J., & Schuberth, F.: Using confirmatory composite analysis to assess emergent variables in business research. Journal of Business Research. **120**, 147--156 (2020)
16. Liengaard, B., Sharma, P.N., Hult, T.M.H., Jensen, M.B., SarstedtHair, M.J.F. and Ringle, C.M.: Prediction: Coveted, yet Forsaken? Introducing a Cross-Validated Predictive Ability Test in Partial Least Squares Path Modeling, Decision Sciences, in press (2020)
17. Muller, D., Judd, C. M., & Yzerbyt, V. Y.: When moderation is mediated and mediation is moderated. Journal of personality and social psychology. **89**(6), 852--863 (2005)
18. Pakpahan, E., Hoffmann, R., & Kröger, H.: Statistical methods for causal analysis in life course research: an illustration of a cross-lagged structural equation model, a latent growth model, and an autoregressive latent trajectories model. International Journal of Social Research Methodology. **20**(1), 1--19 (2017)
19. Sarstedt, M., Ringle, C. M., & Hair, J. F.: Treating Unobserved Heterogeneity in PLS-SEM: A Multi-method Approach. In Partial Least Squares Path Modeling, pp. 197–217. Cham: Springer International Publishing (2017)
20. Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P.: Estimation issues with PLS and CBSEM: where the bias lies!. Journal of Business Research. **69**(10), 3998--4010 (2016)
21. Sharma, P. N., Shmueli, G., Sarstedt, M., Danks, N., & Ray, S.: Prediction-oriented model selection in partial least squares path modeling. Decision Sciences. **52**(3), 567--607 (2021)
22. Sharma, P., Sarstedt, M., Shmueli, G., Kim, K. H., & Thiele, K. O.: PLS-based model selection: The role of alternative explanations in information systems research. Journal of the Association for Information Systems. **20**(4), 4 (2019)
23. Shmueli, G., Ray, S., Estrada, J. M. V., & Chatla, S. B.: The elephant in the room: Predictive performance of PLS models. Journal of Business Research. **69**(10), 4552--4564 (2016)
24. Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J. H., Ting, H., Vaithilingam, S., & Ringle, C. M.: Predictive model assessment in PLS-SEM: guidelines for using PLSpredict. European Journal of Marketing. **53**, 2322--2347 (2019)
25. Wagenmakers, E. J., & Farrell, S.: AIC model selection using Akaike weights. Psychonomic bulletin & review. **11**(1), 192--196 (2004)

PLS-SEM with CCA for football goalkeeper's performance indicators

PLS-SEM con CCA per indicatori di performance dei portieri di calcio

Mattia Cefis and Maurizio Carpita

Abstract Today, PLS-SEM is a very trend-topic, while football analytics is an emerging field of research; by this contribution we aim to give a new approach in the evaluation of football goalkeepers' performance given from the EA Sports experts by data available on the Kaggle data science platform. For this purpose, we adopt an innovative confirmatory composite analysis (CCA) to validate and evaluate a second-order formative-formative PLS-SEM model. After its validation, we compare this new indicator with a benchmark (the EA *overall*) and respectively goalkeepers' wage and market value. The final goal is to prove the CCA approach on a real case study and to suggest an original performance indicator for helping coaches and scouting staff of professional teams to take strategic decisions.

Abstract *Al giorno d'oggi, il PLS-SEM è un argomento di tendenza mentre la statistica applicata al mondo del calcio è un campo di ricerca emergente; con questo lavoro abbiamo l'obiettivo di fornire un nuovo approccio nella valutazione della performance dei portieri partendo da quella già offerta dagli esperti di EA Sports grazie a dati disponibili sulla piattaforma Kaggle. Per quest'obiettivo adotteremo un nuovo approccio, la confirmatory composite analysis (CCA) per validare e valutare un modello formativo-formativo PLS-SEM di secondo ordine. Dopo la sua validazione confronteremo questo nuovo indicatore con l'EA overall, con il salario ed il valore di mercato dei portieri. L'obiettivo finale è quello di verificare la CCA su un caso di studio e proporre un indicatore originale di performance per aiutare allenatori e l'area scouting di una società calcistica a prendere decisioni strategiche.*

Key words: Composite indicators, CCA, PLS-SEM, Football KPI

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

The latest developments in sports research are moving towards a data-driven approach. In particular, focused on football (i.e. soccer for Americans), players' performance evaluation is becoming a strategic key for football coaches and policy makers. The majority of papers on performance evaluation are focused just on movement players (i.e. defenders, midfielders and forwards, [5]): by this research we want to focalize attention on a singular role, the goalkeeper. We know that goalkeepers' performance on the soccer field has been measured and described by Electronic Arts (EA)[1] experts. In their opinion, goalkeepers' performance can be thought as a multidimensional construct made up of 7 performance composite indicators (i.e. the same 6 used for movement players plus a specific one for goalkeepers, due to their singular function), each one made up of several specific skills, which combined form an *overall* index that sums up the performance; then, a statistical support is required [1, 3]. In this paper, our goal is to propose the use of an innovative confirmatory composite analysis (CCA) to validate a second order formative-formative Partial Least Squares - Structural Equation Model (PLS-SEM) model starting from the data provided by the Kaggle data science platform, in order to build a new composite indicator specific for the goalkeepers and to compare it with the well-known EA *overall* and other proxies (i.e. goalkeepers' wage and market value), in order to verify the CCA procedure and give a significant statistics support to the experts' opinion both.

2 Literature overview and data employed

Existing literature focused on players' performance [1, 3] includes different approaches: for example Carpita [2] adopted an unsupervised method to classify different area of performance, Cefis and Carpita [5] already proposed a PLS-SEM model considering movement roles. The aim of this research is to focalize attention on the evaluation of goalkeepers' performance, exploring key performance indices (KPIs), in order to evaluate some different strategic skills.

For this application has been used data from EA experts and available on the famous Kaggle data science platform by Leone[2]; in particular, we will focus on all goalkeepers' stats from the top 5 European Leagues (e.g., Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). This dataset contains 31 variables (e.g. KPIs), with periodic players' performance on a 0-100 scale with respect to different abilities, classified by *sofifa* experts into 6 latent traits: *attacking, skill, movement, power, mentality* and *goalkeeper features*; note that, after a preliminary check, we did not take into account the *defending* block for this model, since its skills are strictly related with movement players (i.e. *mark-*

[1] www.easports.com

[2] www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

ing, sliding tackles and *standing tackle*). For our purpose we have chosen to take into account data relying the beginning of the season 2019/2020, so the dataset was composed by stats about 331 goalkeepers.

2.1 The second-order PLS-SEM model and the CCA approach

PLS-SEM [15], also called PLS-PM, is a tool that offers a valid alternative to the well-known covariance-based model [10]. Its goal is to measure causality relation between concepts (e.g. latent variables), starting from some manifest variables, by an exploratory approach: the explained variance of the endogenous latent variables is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression. Additionally, PLS-SEM does not require any preliminary assumptions for the data, so it's called a soft-modelling technique. In our framework, PLS-SEM estimates simultaneously two models:

- Measurement (outer) model $\Rightarrow$ links MVs to their LVs. Each block of MVs $\mathbf{X}_g$, $g = 1,...,G$ (with $G = 6$ in our case) must contain at least one MV and this relation can be treated in two ways: reflective (where the MVs are the effects of their own LV) and formative (where the MVs are the causes of their own LV). In our framework we will assume a formative structure for the outer model where each LV ξ_g is considered to be formed by its own KPIs following a multiple regression:
$$\xi_g = \mathbf{X}_g \mathbf{w}_g + \delta_g \tag{1}$$
where $\mathbf{w}_g$ is the vector of the outer regression weights (estimated by OLS) and δ_g is the vector of error terms.
- Structural (inner) model $\Rightarrow$ by this model LVs are divided into two groups: exogenous and endogenous. The first one does not have any predecessor in the path diagram, the rest are endogenous. For the *j*-th endogenous variable in the model, the linear equation of its own structural model is:
$$\xi_j = \beta_0 + \sum_{r=1}^{R} \beta_{rj} \xi_r \ + \zeta_j \tag{2}$$
where R is the number of exogenous LVs that affect the endogenous one and β_{rj} is so called path coefficient, a linkage between the *r*-th exogenous LV and the *j*-th endogenous LV and ζ_j is the error term.

Unlike the psychological models, that usually assume reflective relation between concepts [6], here we will adopt, following some experts suggestion[3], a formative-formative approach (for measurement and structural model both), since our latent traits of performance are not directly measurable and at the same are "made" by their own KPIs; starting from this assumption, we will assume a PLS-SEM with

[3] https://www.fifauteam.com/fifa-19-attributes-guide/

second-order construct, also known as hierarchical model [12]. In this framework we can include LVs that represent a "higher-order" of abstraction (i.e. higher order construct, HOC). In fact, for our purpose, we will assume goalkeepers' composite performance as extra-latent construct of higher (second) order, influenced directly from the others 6 exogenous ones. Since this HOC is without any apparent MVs, literature suggested us a technique in order to modelling this framework: a two-step approach [12]. In the first step we compute by Principal Component Analysis the scores of the lower-order constructs (e.g. the first principal component -I PC- of each one), while in the second one we can apply the classical PLS-SEM using the computed scores as MVs for the endogenous.

In order to validate our model we will apply an innovative CCA approach, that is a systematic methodological process for confirming measurements models in PLS-SEM [8]: it is explanatory and confirmatory both. Already Ciavolino et al. [6] adopted this approach for confirming a reflective-reflective psychological framework. We must take in mind that formative indicators cannot be evaluated at the same manner of reflective ones: for this reason, the CCA suggests different steps for what concerning the measurement model evaluation:

1. *Convergent validity*: it is based on the size of the path coefficient between two constructs. Hair et al. [13] recommands that the larger the size of the coefficient, the stronger is the indication of convergent validity.
2. *Indicator multicollinearity*: it is suggested to adopt the well-known variance inflation factor (VIF); if this index is lower or equal to 3.0, then multicollinearity is not a problem.
3. *Significance of indicator weights*: it is evaluated by a bootstrapping test with 5% significance (it is suggested to set $\alpha = 10\%$ for small sample sizes).
4. *Contribution of indicators loadings*: a loading is considered important in forming the construct when it is greater or equals to 0.50 and statistically significant (by a bootstrapping test).
5. *Predictive validity*: this last step assesses the extent to which a construct score predicts scores on some criterion measure. It involves using the construct score to predict the score of a criterion variable that is collected at a later point in time.

For this project the R software packages csem [11] and seminr [14] have been used; we carried out a bootstrap validation (i.e. 1000 resampling) for the model in order to assess the path significance. In the next section, the results are shown.

3 Results and discussion

After CCA validation, the final model is showed in Fig. 1: we removed 5 KPIs with non significant weights and 2 due to their multicollinearity problem; the others steps of CCA hold. We can see how *power, mentality* and *GK_Features* (as we expected) have the strongest impact on the macro-composite indicator (i.e. beta coefficients significant and greater than 0.20 for the structural model). It's interesting to note

how for each LV the strongest MV (i.e. with highest weight) is a typical variable strictly related with the goalkeepers ability [9]: *long passing* for *skill*, *reaction* for *movement*, *shot power* for *power*, *composure* for *mentality*, *short passing* for *attacking* and *GK features* for global *GK performance*. About the outer model, for the lower order constructs, the KPIs with VIF greater than 3.0 are *Diving* and *Kicking*, from the *GK features* construct: they are respectively 4.10 and 3.67, but we decided to keep them since their VIF is not too distant from the cut-off and because a formative indicator should never be eliminated based solely on statistical criteria [8]. The model has also a good fit (GoF index = 0.72).

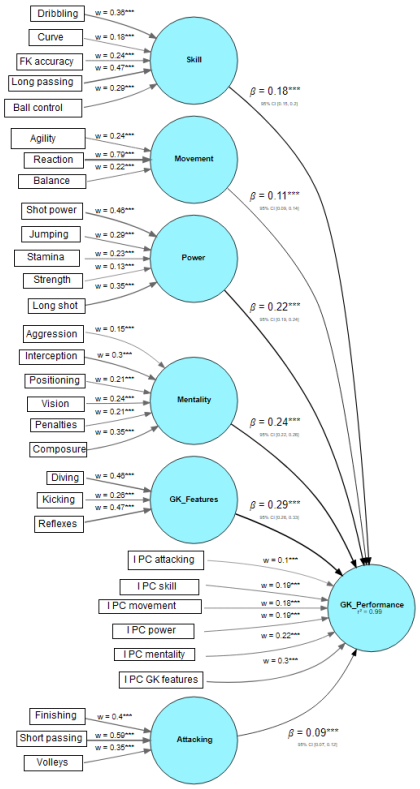


Fig. 1 PLS-SEM GK performance model.

In order to check the predictive validity, we compared our scores with some criterion measure (Tab. 1) relying the beginning of the successful season (2020/2021), such as the EA *overall*, wage and players' market value, with interesting results: all medium-low correlations, but significant (no one CI 95% contains the zero), the highest between our indicator (referred the season 19/20) and the EA *overall* of the season 20/21.

Table 1 Correlations of the GK Performance Indicators with three criterion variables

	GK performance 19/20	CI 95%
EA *overall* 20/21	0.441	$[0.329 - 0.541]$
Wage 20/21	0.309	$[0.185 - 0.422]$
Market Value 20/21	0.258	$[0.131 - 0.376]$

Finally, this model seems to provide comforting results, and at this point for future projects it could be interesting to integrate it in some predictive modelling, such as the expected goal model used in football analytics [7], or maybe to compare results with different higher order approaches [4].

References

1. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the kaggle european soccer database. Statistical Modelling **19**(1), 74–101 (2019)
2. Carpita, M., Ciavolino, E., Pasca, P.: Players' role-based performance composite indicators of soccer teams: A statistical perspective. Social Indicators Research **156**(2), 815–830 (2021)
3. Carpita, M., Golia, S.: Discovering associations between players' performance indicators and matches' results in the european soccer leagues. Journal of Applied Statistics **48**(9), 1696–1711 (2021)
4. Cataldo, R., Grassia, M.G., Lauro, N.C., Marino, M.: Developments in higher-order pls-pm for the building of a system of composite indicators. Quality & Quantity **51**(2), 657–674 (2017)
5. Cefis, M., Carpita, M.: Football analytics: a higher-order pls-sem approach to evaluate players' performance. Book of Short Papers SIS 2021 pp. 508–513 (2021)
6. Ciavolino, E., Ferrante, L., Sternativo, G.A., Cheah, J.H., Rollo, S., Marinaci, T., Venuleo, C.: A confirmatory composite analysis for the italian validation of the interactions anxiousness scale: a higher-order version. Behaviormetrika pp. 1–24 (2021)
7. Green, S.: Assessing the performance of premier leauge goalscorers. OptaPro Blog (2012). URL http://www.optasportspro.com/about/optaproblog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/
8. Hair Jr, J.F., Howard, M.C., Nitzl, C.: Assessing measurement model quality in pls-sem using confirmatory composite analysis. Journal of Business Research **109**, 101–110 (2020)
9. Hughes, M.D., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., Duschesne, C.: Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position (2012)
10. Jöreskog, K.G.: Structural analysis of covariance and correlation matrices. Psychometrika **43**(4), 443–477 (1978)
11. Mehmetoglu, M., Venturini, S.: Structural equation modelling with partial least squares using Stata and R. CRC Press (2021)
12. Sanchez, G.: Pls path modeling with r. Berkeley: Trowchez Editions **383**, 2013 (2013)
13. Sarstedt, M., Ringle, C.M., Hair, J.F.: Partial least squares structural equation modeling. Handbook of market research **26**(1), 1–40 (2017)
14. Shmueli, G., Ray, S., Estrada, J.M.V., Chatla, S.B.: The elephant in the room: Predictive performance of pls models. Journal of Business Research **69**(10), 4552–4564 (2016)
15. Wold, H.: Encyclopedia of statistical sciences. Partial least squares. Wiley, New York pp. 581–591 (1985)

Text-mining and PLS-SEM combination to measure food satisfaction with Google Review: When the gut (re)counts!

Combinare Text-mining e PLS-SEM per misurare la soddisfazione alimentare con Google Review: Quando la pancia (rac)conta!

Paola Pasca and Michelangelo Misuraca and Alessia Meloni and Enrico Ciavolino

Abstract This preliminary study aims to identify customer satisfaction through the analysis of natural language used in restaurant Google Reviews. Through a web scraping procedure, restaurant reviews of the Italian provinces, from 2018 to 2020, were collected. Italian EMotion lexicon (ItEM), a high-coverage emotion lexicon developed for the Italian language, based on Plutchik's taxonomy, was used to isolate the psychological and emotional aspects of customers' perception. Results show the existence of different nuances in the emotional perception associated with restaurants that may account for overall customer satisfaction.

Abstract *Il presente studio preliminare si propone di identificare la customer satisfaction attraverso l'analisi del linguaggio naturale utilizzato nelle recensioni dei ristoranti su Google Reviews. Attraverso una procedura di web scraping sono state raccolte le recensioni dei ristoranti delle province italiane, dal 2018 al 2020. Per isolare gli aspetti psicologici ed emotivi della percezione è stato utilizzato l'Italian EMotion lexicon (ItEM), un lessico delle emozioni ad alta copertura, sviluppato per la lingua italiana, basato sulla tassonomia di Plutchik. I risultati mostrano*

Paola Pasca
Department of History, Society and Human Studies
University of Salento (Lecce, Puglia, Italy)
e-mail: paola.pasca@unisalento.it

Michelangelo Misuraca
Department of Business Administration and Law
University of Calabria
e-mail: michelangelo.misuraca@unical.it

Alessia Meloni
Department of History, Society and Human Studies
University of Salento (Lecce, Puglia, Italy)
e-mail: alessia.meloni1996@gmail.com

Enrico Ciavolino
Department of History, Society and Human Studies
University of Salento (Lecce, Puglia, Italy)
e-mail: enrico.ciavolino@unisalento.it

l'esistenza di diverse sfumature nella percezione emotiva associata ai ristoranti che potrebbe render conto della soddisfazione generale dei clienti.

Key words: Text-mining, PLS-SEM, ItEM, Customer Satisfaction

1 Introduction

Restaurant business has an enormous importance in the economy of every country. More specifically, in Italy in 2019 the restaurant field had a business of almost 86 billions of euro [1], engraving for about 5% point on the national PIL [1].

Recently, the importance of the customer in the service industry has been increasingly recognized. One of the most important objectives in this area is customer satisfaction, retention and loyalty. Customer satisfaction, in fact, emerges as an important aspect both for improving the quality of services and for increasing profit [2]. In addition, it is associated with other aspects, such as customer loyalty, as demonstrated by a number of studies [3, 4].

Works, such as [5], [6] and [7] examined customer satisfaction in restaurant business, while others focused more specifically on customer satisfaction models such as the SCSB model [8], the ACSI model [9] and the ECSI model [10]. The different theoretical models and their evolutions define and consider the central elements of customer satisfaction: for instance, the ECSI model [10] considers Image, Expectation, Hardware and Software. On the other hand, the PROSERV model [11] draws a distinction between the affective dimension (Experience) and the outcome (Value) which in turn determines the constructs of Utility, Co-Construction, Devices, Front-Office and Process. This preliminary study aims to capture customer satisfaction from natural language processing of Google reviews: considering that emotional aspects of the food service experience can be captured by natural language processing [12, 13] and the restaurant ratings (1 to 5 stars) as an expression of general satisfaction, it becomes possible to hypothesize an association between the two. More specifically, it could be possible to imagine that language-derived psychological dimensions may account for restaurants evaluation.

2 Method

2.1 Sample

As [14] suggest, web scraping is a methodology with great potential for psychological research. In fact, it allows gathering large amounts of data from forums, social media and other websites, in a fully automated manner [15]. In this preliminary study data were collected through web scraping from Google Maps review.

In choosing this platform the authors were guided by several choices: first, it is the most used platform in the world to share opinions and reviews; second, it is the most used in finding new restaurants; moreover, it contains an enormous amount of reviews and is considered a reliable and valid platform. To automate the data collection a Python script has been used: given the name of a city as input, the script extracts name and address of all the restaurants in the city, along with the names of the reviewers and their ratings (stars).

For the purpose of this preliminary study, reviews from 2018 to 2020 related to Italian provinces were considered. At the end of data collection, the dataset included more than 4 million reviews. Table 1 shows the distribution of data per region.

Table 1 Distribution of reviews per region

Region	Cities	Reviews	Reviewer	Restaurants
Abruzzo	4	874	758	308
Aosta Valley	1	333	314	77
Apulia	8	4371	3731	1013
Basilicata	2	1139	1001	237
Calabria	5	1052	855	377
Campania	5	7052	5794	921
Emilia-Romagna	10	12246	9332	2381
Friuli Venezia Giulia	4	2151	1791	504
Lazio	5	19088	16092	2898
Liguria	4	3515	2999	703
Lombardy	12	14766	12289	2245
Marche	6	1438	1251	448
Molise	2	189	160	86
Piedmont	7	6241	4885	1638
Sardinia	5	4319	3572	845
Sicily	9	8898	7236	1461
Trentino-South Tyrol	2	1065	932	212
Tuscany	11	17590	14685	2273
Umbria	2	1126	988	292
Veneto	7	11763	10379	1300
Total	111	119216	99044	20219

In order to reduce the dataset size, reviews have been aggregated based on the restaurant they belonged to, region and province, for a total of 19313 restaurants each with their own text, metadata and average rating.

2.2 Text Analysis

Text analysis or text mining is a powerful method which could be used to extract information from written text [16]. It can be divided into two main approaches: on the one hand, a top-down or dictionary-based one, in which words are already cat-

egorized as belonging to certain emotional or psychological categories; and bottom up approaches, which start from the words in order to grasp themes in the texts (e.g. topic modeling, dimension reduction techniques) [13]. However, more and more attempts are being made to create dictionaries consistent with theory, through automated procedures: one example is ItEM [17], an Italian lexicon of emotional words reflecting Plutchik's taxonomy [18], which conceives emotions as spread out over a circle, where the emotion intensity decreases from the inside to the outside. (see Fig. 1).

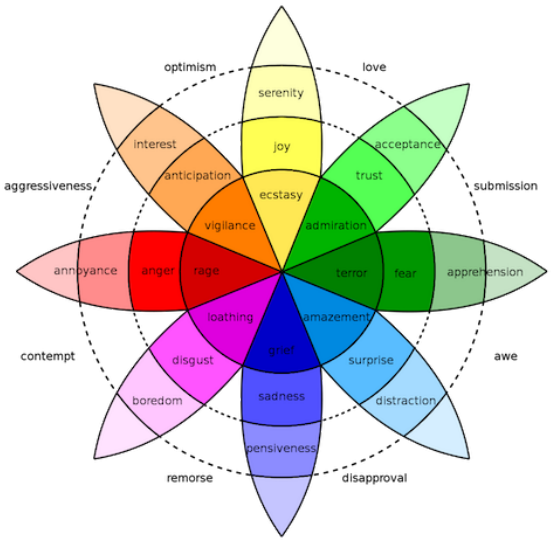


Fig. 1 Plutchik's wheel of emotions

The initial dataset was turned into a Document Term Matrix (DTM), while the ItEM tokens referred to the basic emotions, intermediate intensity (*joy, trust, fear, surprise, sadness, disgust, anger, anticipation*) were used to isolate words of interest within the reviews.

3 Results

In an effort to summarize words in a meaningful way, a Principal Component Analysis (PCA) has been performed on the words belonging to each of the emotional categories [19]. Table 2 shows loadings > 0.2 for the first 5 components of the emotional category *joy*:

Table 2 PCA of the *joy* emotion [a]

	Component				
	1	2	3	4	5
condividere	0.594	-	-	-	-
condiviso	0.540	-	-	-	-
delizia	0.376	-	-	-	-
condivisi	0.358	-	-	-	-
vivace	0.327	-	-	-	-
condividendo	0.277	-	-	-	-
contento	0.266	-	-	-	-
contenti	0.262	-	-	-	-
gioia	0.243	-	-	-	-
eccitato	0.241	-	-	-	-
condivisa	0.231	-	-	-	-
piacere	0.217	-	-	-	-
sorriso	0.205	-	-	-	-
condividi	0.205	-	-	-	-
giocare	-	0.738	-	-	-
giochi	-	0.711	-	-	-
gioco	-	0.611	-	-	-
giocando	-	0.328	-	-	-
cantare	-	-	0.511	-	-
allegria	-	-	0.424	-	-
canta	-	-	0.404	-	-
allegro	-	-	0.324	-	-
cantante	-	-	0.323	-	-
cantanti	-	-	0.316	-	-
allegra	-	-	0.314	-	-
cantava	-	-	0.280	-	-
risate	-	-	0.264	-	-
cantato	-	-	0.249	-	-
cantando	-	-	0.238	-	-
ritroverete	-	-	0.219	-	-
colorato	-	-	-	0.691	-
colorati	-	-	-	0.657	-
colorate	-	-	-	0.503	-
colorata	-	-	-	0.311	-
vincente	-	-	-	0.232	-
sole	-	-	-	-	0.537
bellezze	-	-	-	-	0.475
bellezza	-	-	-	-	0.461
partecipanti	-	-	-	-	0.368
giocano	-	-	-	-	0.247
vince	-	-	-	-	0.218
trionfo	-	-	-	-	0.210

[a] As the English language tends to use the same word to indicate terms of different types (e.g. adjectives, verbs), the words listed in the table are the native Italian ones.

Even considering words belonging to the same basic emotion, results show how different aspects of a positively connoted experience emerge: the first dimension defining *conviviality* (e.g. *sharing, delight, lively*), the second *playfulness* (e.g., *playing, play*), the third *entertainment and participation* (e.g. *sing, cheerfulness, laughter*), the fourth *visual appeal* (e.g. *colorful*), the fifth *competitions* (e.g. *game, participants, win, triumph*). Based on this and other basic emotion results, a PLS-SEM model will be formalized that links the structure of the 8 dimensions and their subdimensions to restaurants ratings.

References

1. Sbraga L., Romana Erba G.: Ristorazione - Rapporto Annuale 2019. FIPE, Federazione Italiana Pubblici Esercizi (2019) https://www.fipe.it/centro-studi/news-centro-studi/item/6817-ristorazione-2019.html
2. Anderson, E. W., Fornell, C., Lehmann, D. R.: Customer satisfaction, market share, and profitability: Findings from Sweden. J. Marketing **58**(3), 53–66. SAGE Publications: Los Angeles, CA (1994) doi: 10.2307/1252310
3. Bowen, J. T., Chen, S. L.: The relationship between customer loyalty and customer satisfaction. Int. J. Contemp. Hosp. Manag. **13**, 213–217. MCB UP Ltd (2001) doi: 10.1108/09596110110395893
4. Ciavolino, E., Lagetto, G., Montinari, A., Al-Nasser, Amjad, D., Al-Omari, A. I., Zaterini, M. J., Salvatore, S.: Customer satisfaction and service domains: a further development of PROSERV Qual Quant **54**, 1429–1444. Springer (2019) doi: 10.1007/s11135-019-00888-4
5. Babin, B. J., Lee, Y. K., Kim, E. J., Griffin, M.: Modeling consumer satisfaction and word-of-mouth: restaurant patronage in Korea. J. Serv. Mark., **19**, 133–139 (2005) doi: 10.1108/08876040510596803
6. Andaleeb, S. S., Conway, C.: Customer satisfaction in the restaurant industry: an examination of the transaction-specific model. J. Serv. Mark., **20**, 3–11 (2006) doi: 10.1108/08876040610646536
7. Hwang, J., Zhao, J.: Factors influencing customer satisfaction or dissatisfaction in the restaurant business using AnswerTree methodology. J. Qual. Assur. Hosp. Tour., **11**(2), 93–110 (2010)
8. Fornell, C.: A national customer satisfaction barometer: The Swedish experience. J. Marketing, **56**(1), 6–21 (1992) doi: 10.2307/1252129
9. Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., Bryant, B. E.: The American customer satisfaction index: nature, purpose, and findings. J. Marketing, **60**(4), 7–18 (1996)
10. ECSI Technical Committee: European customer satisfaction index: foundation and structure for harmonized national pilot projects. Report prepared for the ECSI Steering Committee (1998)
11. Ciavolino, E., Salvatore, S., Mossi, P., Vernai, M.: Quality and prosumership. proserv: a new tool for measuring the customer satisfaction. Int. J. Bus. Soc., **18**(3), 409–426. (2017)
12. Passaro, L. C., Lenci, A.: Evaluating context selection strategies to build emotive vector space models. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2185–2191 (2016)
13. Kennedy, B., Ashokkumar, A., Boyd, R. L., Dehghani, M.: Text Analysis for Psychology: Methods, Principles, and Practices. PsyArXiv (2021) https://doi.org/10.31234/osf.io/h2b8t
14. Landers, R. N., Brusso, R. C., Cavanaugh, K. J., Collmus, A. B.: A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychol. Methods, **21**(4), 475. (2016)
15. Boyd, R. L., Pasca, P., Lanning, K.:The personality panorama: Conceptualizing personality through big behavioural data. Eur. J. Pers., **34**(5), 599–612. (2020)
16. Settanni, M., Marengo, D.: Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. Front. Psychol., **6** 1045. (2015) doi: 10.3389/fpsyg.2015.01045
17. Passaro, L., Pollacci, L., Lenci, A.: ItEM: A vector space model to bootstrap an italian emotive lexicon. In Second Italian Conference on Computational Linguistics CLiC-it. 215–220. Academia University Press. (2015)
18. Plutchik, R. E., Conte, H. R.: Circumplex models of personality and emotions. American Psychological Association. Washington, DC (1997) ISBN: 978-1-55798-380-0
19. Jolliffe, I.: Principal component analysis. In Everitt, B., Howell, D.: Encyclopedia of statistics in behavioral science. Wiley (2005) doi: 10.1002/0470013192

Session of solicited contributes SS8 – *Applications of non standard statistical tools to real-life*
Organizer and Chair: Antonio D'Ambrosio

A robust strategy for building a financial portfolio

Una strategia robusta per la costruzione di portafogli finanziari

Carmela Iorio and Giuseppe Pandolfo

Abstract The mean-variance portfolio constitutes the milestone of the modern portfolio theory. The mean-variance model relies on two fundamental assumptions. First, a rational investor maximizes, over a single period, the expected return of an asset for a given level of risk, which is measured by the variance of stock returns themselves. Second, the random returns are normally distributed. In reality, it is well-known that the time series of returns have heavier tails and a higher peak than in a normal distribution. In this paper, we propose the application of statistical weighted depth functions as an alternative non-parametric tool. The aim is to build a robust mean-variance model within the standard portfolio selection framework. Real data are used to investigate the performances of the proposed approach.

Abstract *La pietra miliare della moderna teoria di portafoglio è costituita dal modello media-varianza. Tale modello si fonda su alcune ipotesi. In primo luogo, un investitore razionale massimizza, in un singolo periodo, il rendimento atteso di un'attività per un dato livello di volatilità che è misurato dalla varianza dei rendimenti delle azioni stesse. In secondo luogo, i rendimenti casuali sono distribuiti normalmente. Tuttavia, è noto che le serie storiche dei rendimenti sono caratterizzate da code più "pesanti" rispetto a quelle di una distribuzione normale. In questo paper, si propone l'uso di un metodo di stima non parametrico per la costruzione di un portafoglio media-varianza "robusto". Le prestazioni dell'approccio proposto sono studiate con un'applicazione a un set di dati reale.*

Key words: Time Series, Portfolio selection, Mean-variance portfolio, Non-parametric estimation method.

Carmela Iorio
Department of Economics and Statistics, University of Naples Federico II e-mail: carmela.iorio@unina.it

Giuseppe Pandolfo
Department of Economics and Statistics, University of Naples Federico II e-mail: giuseppe.pandolfo@unina.it

1 General Framework

Investors allocate capital across a set of stocks in order to maximize the return and minimize the risk. Portfolio theory describes how investors should allocate their wealth. The process of efficiently allocating wealth among assets has a long history in literature. The pioneer work by [3] constitutes the milestone of Modern Portfolio Theory (MPT). The Markowitz model, known as Mean-Variance (MV) portfolio, aims to build an optimized portfolio by selecting stocks having the highest expected return for given level of risk, which is measured by standard deviation of the assets returns. Hence, assets covariances reflect the importance of diversification in mitigating selection risks. Given a set of $N \geq 2$ financial assets, the mean value, i.e. the expected return, and the variance of a portfolio are given by:

$$\mu_N = E(R_N) = \sum_{n=1}^{N} w_n E(R_n)$$

$$\sigma_N^2 = var(R_N) = \sum_{n=1}^{N} w_n^2 \sigma_n^2 + \sum_{n=1}^{N} \sum_{m=1}^{N} w_n w_m \sigma_{nm},$$

where σ_{nm} is the covariance between the return for assets n and m, $\sigma_m = \sqrt{\sigma_m^2}$ and $\sigma_n = \sqrt{\sigma_n^2}$ are the standard deviations of R_m and R_n, respectively.
The covariance between the two assets is computed as $\sigma_{nm} = \sigma_m \sigma_n \rho_{nm}$, where ρ_{nm} indicates the correlation between the returns of assets n and m. Firstly, the weights must be found. Then, for a given level of expected return R^*, the portfolio with minimum variance is selected. The MV portfolio optimization problem can be mathematically formulated as the following quadratic problem:

$$\min\ w'\Sigma w \quad s.t. \quad I'w = 1 \quad and \quad w'R = R^*,$$

where $\Sigma = [\sigma_{nm}]_{1 \leq n,m \leq N}$ is the covariance matrix, I is the identity matrix and R is the expected return vector ($N \times 1$). The expected return for the n-th asset in the portfolio is denoted by $\mu_n = E(R_n)$.
Markowitz model has become quite popular in the financial industry, mainly due to the natural and intuitive formulation. However, it has been criticized over time due to its assumptions. It is known that the asset returns are characterized by heavier tails and a higher peak than in a normal distribution [6]. Of course, under normality the solutions obtained through the Maximum Likelihood Estimation (MLE) are the most efficient. Neverthless, when deviations from the Normal distribution occur the resulting solutions may be heavily not stable and the bias of these estimators can be very large. Slight changes in covariance matrix can significantly change the portfolio allocations. To overcome this drawback, we propose to use a *robust* approaches by exploiting the notion of data depth function. This model risk problem is known as a problem of statistical robustness [4]. In the following, we first recall the definition of data depth function and introduce the robust estimates of the mean and covariance matrix by using a depth weighted function, then the results of an application to a real data set are offered to the reader.

2 The proposal

The impulse to the use and development of non-parametric methods in the last decades is mainly due to untenable assumptions of classical parametric approach. The concept of data depth is an important non-parametric tool in multivariate data analysis. The notion of data depth was introduced by [7] as a graphical tool for bivariate data. Then, [2] extended data depth concept to the multivariate case. Data depth concept leads to a natural center-outward ordering of sample points in multivariate data sets as well, and extends univariate concepts based on order to higher dimensions. The depth of a point relative to a given data set measures how central that point is with respect to the distribution (the lower the depth of a point is, the more outer this point is). Thus, the deepest point is a multivariate location parameter. Several approaches can be considered to define a depth function [10]. In this paper, the focus is put on a distance-based approach. This means that the distance/outlyingness, and hence the depth, of a point x from the points $x_1,\ldots,x_r$ can be measured by a non-negative distance function. L^p depth measures in some sense the mean outlyingness of a point $x \in \mathbb{R}^d$ via the L^p norm $\|\cdot\|_p$ ($p > 0$):

$$L^pD(x,F) = \frac{1}{1+E\|x-X\|_p},$$

where $X \sim F$. For $p = 2$ the Euclidean norm is obtained.
The L^p depth vanishes at infinity, and is maximum at the point $x \in \mathbb{R}^d$ that minimizes $E\|x-X\|_p$. Monotonicity with respect to the deepest point, convexity and compactness of the central regions derive from the triangle inequality. $L^pD(x,F)$ generally does not satisfy the affine invariance property. On the other hand, for $p = 2$ it becomes invariant under rigid Euclidean transformations. Different distances with respect to the data are usually treated with equal importance (equally weighted) even if the importance may not be the same for different distances in some cases. An usual way to obtain location and scatter estimators, designed to achieve greater robustness, is down-weighting the more outlying observations. [11] defined the weighted L^p depth as:

$$WL^pD(x,XF) = \frac{1}{1+E[\psi(||x-X||_p)]},$$

where ψ is a weight function assumed to be non-decreasing and continuous on $[0,\infty)$, and $X \sim F$. We obtain the robust estimates of μ and Σ by using a depth weighted function as location and scatter estimator [8, 9, 5] which is, respectively, defined as follows:

$$\hat{\mu}_{WL^pD} = \frac{\sum_{i=1}^n \psi_1\{D(x_i,F)\}x_i}{\sum_{i=1}^n \psi_1\{D(x_i,F)\}}$$

$$\hat{\Sigma}_{WL^pD} = \frac{\sum_{i=1}^n \psi_2\{D(x_i,F)\}(x_i-\hat{\mu}_{WL^pD})(x_i-\hat{\mu}_{WL^pD})'}{\sum_{i=1}^n \psi_2\{D(x_i)\}}$$

where ψ_1 and ψ_2 are non-decreasing, non-negative smooth weight functions which may not be the same. The choice of the weight function have a great impact on

the relative efficiency and robustness [11]. As suggested by [12], to provide a good balance between efficiency and robustness, we adopt:

$$\psi_j = \frac{\exp\left[-k\left\{1-(r/c)^{2j}\right\}^{2j}\right]-\exp(-k)}{1-\exp(-k)} I_{(0<r<c)} + I_{(c<r<1)},$$

where $j = 1, 2$, k controls the degree of approximation, and $0 < c < 1$ is the median of the depth function. $I(\cdot)$ denotes the indicator function, while $0 \leq r \leq 1$ indicates the empirical depth values.
The functions ψ_j assign weight 1 to the half of the points with higher depth values, while low weights are given to the other half of points with lower depth values. The weighted L^p depth estimators assign small weights to specific observations according to their "abnormal" influence on the estimation by actually considering the whole amount of data simultaneously. We want to highlight that the weighted L^p depth estimators adopted here do not remove information from the sample by "truncating" some data. In this way, it is possible to get more insights about the structure of the model, which is, instead, often not captured by classical estimation methods and those based on trimming or Winsorization.

3 Experimental results

In this section, we show an application of our proposal on real financial time series. The data set consists of 29 assets of DAX 30, traded on Frankfurt Stock Exchange (we did not consider the *Vonovia SE* component stock because it has been listed only since 2013). The DAX is a blue chip stock market index focusing on the large-cap sector of the German market. The data are provided by *yahoo.finance.com* and were monthly collected from December 2003 to February 2018. The time period includes the well-known period of the subprime crisis and the market correction of May 2006. Fig. 1 shows the financial time series of the returns of these assets. The returns are expressed in terms of log-price difference of stocks. As it can be noticed by looking at Fig.1, there are some outlying points corresponding to the negative financial events occurred during the time span. To show the advantages of the proposed estimation technique, we perform an out-of-sample evaluation of portfolio based the weighted L^2 depth-based estimator. The proposed portfolio is compared with different strategies (traditional and robust). Specifically, we consider: the classical Maximum Likelihood Estimator of the mean-variance model (MV) with risk aversion parameter $\lambda = 1$; the M-estimator (M); the minimum S-estimator (S); the minimum covariance determinant estimator (MCD) and the minimum volume ellipsoid (MVE). To compare the performances of the above mentioned methods, we use a rolling horizon procedure like the one proposed by [1]. We used an estimation window length of $\tau = 58$ ($\approx$ 5 years), leaving the last $T - \tau = 112$ months ($\approx$ 9 years) for the out-of-sample evaluation. The out-of-sample performance of each

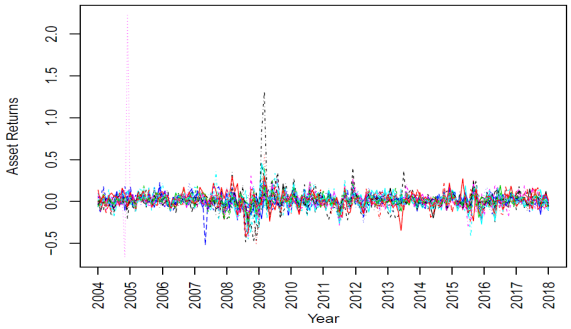


Fig. 1 Financial time series containing the monthly returns of 29 stocks belonging to DAX blue chip stock market index from December 2003 to February 2018)

strategy was evaluated according to variance (σ^2), Sharpe ratio (SR), and portfolio turnover (TO). These measures were computed as follows:

$$(\hat{\sigma}^s)^2 = \sqrt{\frac{1}{T-\tau-1}\sum_{t=\tau}^{T-1}\left(w_t^s r_{t+1} - \hat{\mu}^s\right)^2}, \widehat{SR}^s = \frac{\hat{\mu}^s}{\hat{\sigma}^s},$$

$$TO^s = \frac{1}{T-\tau-1}\sum_{t=\tau}^{T-1}\sum_{j=1}^{N}\left(\left|w_{j,t+1}^s - w_{j,t}^s\right|\right),$$

where the mean excess return is $\hat{\mu}^s = \frac{1}{T-\tau}\sum_{t=\tau}^{T-1} w_t^s r_{t+1}$ with $w_{j,t}^s$ denoting the portfolio weight assigned by the strategy s to the j-th asset at time $t+1$ before rebalancing, and $w_{j,t+1}$ is the desired portfolio weight in the j-th asset at time $t+1$. The results for all the considered techniques are reported in Table 1. It can be noticed that the WL^{2}D-based strategy yields the highest Sharpe ratio (0.339) Also in terms of portfolio turnover, the WL^{2}D-based strategy achieves the best result (0.153), smaller than those achieved by the mean-variance and S-estimator based strategy (0.168 and 0.183, respectively). The highest portfolio turnover was achieved by the M-estimator based strategy.

Table 1 Out-of-sample performance (including p-value for the difference in Sharpe ratios with respect to the mean-variance strategy) of the considered portfolio strategies for the DAX 30.

Statistic	MV	M	S	MCD	MVE	WL^2D
Mean	0.007	0.008	0.009	0.009	0.008	0.009
Variance	0.001	0.001	0.001	0.000	0.001	0.000
Turnover	0.168	0.685	0.183	0.627	0.673	0.153
Sharpe Ratio	0.216	0.243	0.258	0.248	0.234	0.339
P-value	1.000	0.286	0.000	0.174	0.485	0.000

4 Concluding remarks

Portfolios constructed by the classical mean-variance model are sensitive to deviations from the assumption of multivariate normality. To overcome this drawback, we exploited the notion of the weighted L^2 depth function to obtain robust estimates of the mean and covariance matrix of the asset returns. Our proposal has the advantage to be independent of parametric assumptions, and less sensitive to changes in the asset return distribution than traditional technique. We presented a comparison of the proposed estimation technique compared with the classical mean-variance model and four robust techniques through real data. We evaluated the performance of these techniques for 29 component stocks of DAX 30, traded on Frankfurt Stock Exchange between December 2003 and February 2018. Our results indicate that the performance provided by the weighted depth-based procedure performs better than the other strategies. Finally, the proposed mean-variance procedure through data depth is a valuable alternative within portfolio selection theory.

References

1. DeMiguel, V., Nogales, F.J.: Portfolio selection with robust estimation. Operations Research **57**(3), 560–577 (2009)
2. Donoho, D.L., Gasko, M.: Breakdown properties of location estimates based on halfspace depth and projected outlyingness. The Annals of Statistics **20**(4), 1803–1827 (1992)
3. Markowitz, H.: Portfolio selection. The Journal of Finance **7**(1), 77–91 (1952)
4. Perret-Gentil, C., Victoria-Feser, M.P.: Robust mean-variance portfolio selection (2005) Available at SSRN 721509. `https://ssrn.com/abstract=721509orhttp://dx.doi.org/10.2139/ssrn.721509`
5. Serfling, R.: Depth functions in nonparametric multivariate inference. DIMACS Series in Discrete Mathematics and Theoretical Computer Science **72**, 1–16 (2006)
6. Tsay, R.S.: Analysis of financial time series. John Wiley & Sons (2005)
7. Tukey, J.: Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians **2**, 523–531 (1975)
8. Zuo, Y., Cui, H., He, X.: On the stahel-donoho estimator and depth-weighted means of multivariate data. The Annals of Statistics **32**(1), 167–188 (2004)
9. Zuo, Y., Cui, H., Young, D.: Influence function and maximum bias of projection depth based estimators. The Annals of Statistics **32**(1), 189–218 (2004).
10. Zuo, Y., Serfling,R.: General notions of statistical depth function. The Annals of Statistics **28**(2), 461–482 (2000)
11. Zuo, Y., Serfling R.: Robustness of weighted L^2-depth and L^2-median. Allgemeines Statistisches Archiv **88**(2), 215–234 (2000)
12. Zuo, Y., Cui, H.: Depth weighted scatter estimators. The Annals of Statistics **33**(1), 381–413 (2005)

Conditional copula a financial application

La copula condizionata una applicazione finanziaria

Marta Nai Ruscone and Giovanni De Luca

Abstract Understanding the underlying mechanism of influence that are present in financial market is a great challenge. In this work the coditional copula function is presented. In some context, the dependence structure between two variables can be highly influenced by one or more covariates, so it is of interest to know how this dependence structure changes with the value taken by the covariates. An application is carried out to estimate the influence of economic sectors on 46 large companies included in the EUROSTOXX50.

Abstract *Capire il meccanismo che influenza il mercato finanziario è una grande sfida. Il focus di questo lavoro è la funzione copula condizionata. In alcuni contesti, la struttura di dipendenza tra due variabili può essere fortemente influenzata da una o più covariate, quindi è interessante sapere come questa struttura di dipendenza cambia con il valore assunto dalle covariate. Un'applicazione per stimare l'infuenza dei settori economici é presentata per 46 grandi aziende incluse nell'EUROSTOXX 50.*

Key words: Copula function, conditioning, financial returns.

1 Copula

One way used to capture the dependence structure of a multivariate distribution is the copula distribution function. Whenever copula can be defined for any multi-

Marta Nai Ruscone
University of Genova,
e-mail: marta.nairuscone@unige.it

Giovanni De Luca
University of Naples Parthenope,
e-mail: giovanni.deluca@uniparthenope.it.

variate distributions in $\mathbb{R}^d$, we focus on bivariate continuous random vectors for expository purpose.

Let denote $F_{Y_1,Y_2}(y_1,y_2)$ the bivariate cumulative distribution of the pair (Y_1,Y_2) of the random variables Y_1 and Y_2, $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ the marginal c.d.f. of the Y_1 and Y_2, respectively. As show in [4], the joint c.d.f. of (Y_1,Y_2) can be written as:

$$F_{Y_1,Y_2}(y_1,y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2) = C(F_{Y_1}(y_1), F_{Y_2}(y_2))$$

where C is the c.d.f. of the distribution on $[0,1]^2$, with the uniform margins. When variables are continuous, C is unique, and is called the copula of (Y_1,Y_2). Sklar theorem [4] allows to separate the marginal feature and the dependence structure which is represented by the copula. The function C is the c.d.f. of the pair (U,V) where $U = F_{Y_1}(Y_1)$ and $V = F_{Y_2}(Y_2)$, and

$$c(u,v) = \frac{\partial^2 C}{\partial u \partial v}(u,v),$$

is the associated p.d.f. Sklar's theorem proves the existence and the uniqueness of the copula. It also explains how to construct it from the initial distribution. Indeed, for any $0 \leq u,v \leq 1$, the copula is given by

$$C(u,v) = F_{Y_1,Y_2}(F_{Y_1}^{-1}, F_{Y_2}^{-1}),$$

where $F_{Y_1}^{-1}$ and $F_{Y_2}^{-1}$ are the marginal quantile functions. The copula characterizes any nonlinear dependence which is invariant by increasing transformation of either Y_1 and Y_2 . More precisely we have the following : if ϕ and ψ are strictly increasing functions, then (Y_1,Y_2) and $(\phi(Y_1),\psi(Y_2))$ have the same copula.

However, the dependence structure between two variables can be highly influenced by covariates, and it is of interest to know how this dependence structure changes with the value taken by the covariates. This motivates the need for introducing conditional copulas, and the associated conditional Kendall's τ association measure. Conditional copulas have been formally introduced by [2, 3] . They are rather straightforward extensions of the latter concepts, when dealing with conditional distributions. Suppose that the conditional distribution of (Y_1,Y_2) given the values of the covariates fixed at a given level, say $\mathbf{X}$ exist and denote the corresponding conditional joint distribution function by

$$H_X(y_1,y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2 | X = x).$$

If the marginal of H_X denoted as

$$F_{1X}(y_1) = P(Y_1 \leq y_1 | X = x), \qquad F_{2X}(y_2) = P(Y_2 \leq y_2 | X = x)$$

are continuous, then according to Sklar's theorem [1] there exists a unique copula C_X which equals

$$C_X(u_1,u_2) = H_X(F_{1X}^{-1}(u_1), F_{2X}^{-1}(u_2)),$$

where $F_{1X}^{-1}(u) = inf\{y : F_{1X} \geq u\}$ is the conditional quantile function of Y_1 given $X = x$ and F_{2X}^{-1} is the conditional quantile function of Y_2 given $X = x$. The conditional copula C_X fully describes the conditional dependence structure of (Y_1, Y_2) given $X = x$.

The associated conditional Kendall's τ associated measure is given by:

$$\tau_X(u_1, u_2) = \tau(u_1|X, u_2|X).$$

2 Estimating the influence of economic sectors

Categorizing a company into only one industrial sector cannot reflect its whole performance and associated risk. Many listed companies in the stock market belong to conglomerates, conducting their business in different industry sectors; hence, these companies' performance will naturally be influenced by multiple industries. Even if a company only conducts its business in one sector, its performance can still be influenced by other sectors because of the division of labour in modern society.

We have studied the multiple-sector influence on 46 stocks included in the EUROSTOXX50 index (observed in the period 2008/01/02-2019/05/28) making use of the partial dependence methodology.

We use the sector classification from the Global Industry Classification Standard (GICS):
1 Communication services
2 Consumer discretionary
3 Consumer staples
4 Energy
5 Financial
6 Health care
7 Industrials
8 IT
9 Materials
10 Real estate
11 Utilities

Let us define by X, Y and Z the daily returns of three stocks and by M the market index Eurostoxx50 return. The first step is the study of the influence of a stock Z on the pair X and Y. We define the *Influence* quantity

$$d_\tau(X, Y : Z) = \tau(X, Y : M) - \tau(X, Y : M, Z)$$

$\tau(X, Y : M)$ and $\tau(X, Y : M, Z)$ are Kendall's τ estimated from D-Vine copula, where the bivariate copulas are selected among an Elliptical copula (Student's t), and two Archimedean copulas ($BB1$ and $BB7$).

Then we compute the average influence of stock Z on τ between stock X and all the other stocks in the system,

$$d_\tau(X:Z) = \langle d_\tau(X,Y:Z)\rangle_{Y\neq X}$$

where $\langle\rangle$ represents average.

$d_\tau(X:Z)$ approximates the net influence from stock Z to stock X, excluding the influence from the index.

The *average* influence by sector S is given by

$$d_X^S = \frac{1}{N_S}\sum d_\tau(X:Z_S)$$

where X is the investigated stock, Z_S represents the stocks in sector S and N_S is the number of stocks in sector S.

Finally, the average influence is normalized, computing

$$\beta_X^S = \frac{d_X^S}{\sum_S d_X^S}$$

which can be interpreted as the relative influence of sector S on stock X.

In Figures 1-4 we report the β_X^S for four stocks: Axa, Bayer, Airbus, Daimler. We can observe that for stock Axa the Financial, Energy and IT sectors are highy influential. For Bayer and Airbus, we detect the prominent role of a sector, Financial sector and Consumer discretionary sector, respectively. Finally, there are four influential sector for Daimler with approximately equal weight, Financial sector, Energy, Materials and Industrials.

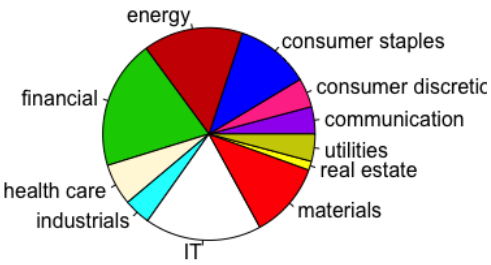


Fig. 1 Relative influence of sectors on Axa.

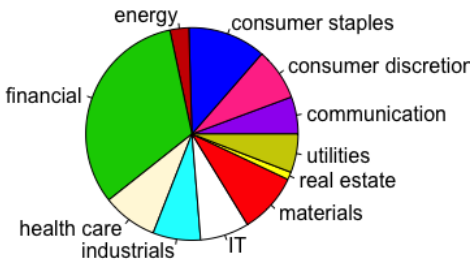


Fig. 2 Relative influence of sectors on Bayer.

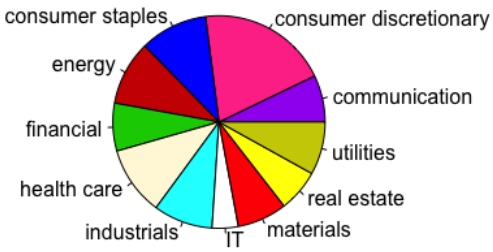


Fig. 3 Relative influence of sectors on Airbus.

3 Conclusions

The dependence structure between two variables can be highly influenced by one or more covariates. Taking into account the conditioning variable can lead to a different perspective of the relationship among financial assets. The approach has been used to estimate the influence of economic sectors on a large sample of companies.

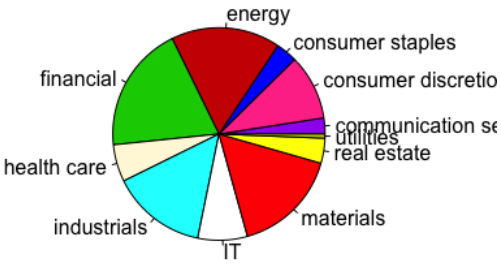


Fig. 4 Relative influence of sectors on Daimler.

References

1. Nelsen, R.: An introduction to copulas. 2nd ed. Springer, New York (2006)
2. Patton, A.: Modelling asymmetric exchange rate dependence, Internat. Econom. Rev. **47**, 527–556 (2006)
3. Patton, A.: Estimation of multivariate models for time series of possibly different lengths, J. Appl. Econometrics **21**, 147–173 (2006)
4. Sklar, A.: Fonctions de repartition a n dimensions et leurs marges, Publ. Inst. Stat. Univ. Paris **8**, 229–231 (1959)

Explaining Student Satisfaction Assessments: A Natural Language Processing Approach

Analisi Della Soddisfazione degli Studenti: Un approccio basato sul Natural Language Processing

Marco Ortu, Luca Frigau and Giulia Contu

Abstract In this study we present an analysis of students' assessments during the last two year during the COVID19 pandemic. We explained the the students' assessments with the teaching using sentiment and emotion analysis on the issues reported by the students. We analyzed 1389 issues extracting positive and negative sentiment and four emotions: joy, sadness, anger and fear. We used these indicators to explain the overall satisfaction with the teaching as measured by end-of-course questionnaire that students are asked to fill in at the end of each attended course. Our models explain from 49.7% to 72.6% of the total variance of the overall satisfaction with the teacheing, and we found that *joy* emotions, when statistically significant (with $\alpha = 5\%$), show a positive influence on the overall satisfaction with the teacher, while *anger* is a warning signal for critical/improvable situations.

In questo studio presentiamo un'analisi delle valutazioni degli studenti durante gli ultimi due anni di pandemia da COVID19. Abbiamo spiegato le valutazioni degli studenti nei confronti dei loro docent utilizzando la sentiment ed emotions analysis sulle segnalazioni riportate dagli studenti. Abbiamo analizzato 1389 segnalazioni estraendo sentimenti positivi e negativi e quattro emozioni: gioia, tristezza, rabbia e paura. Abbiamo utilizzato queste misure di sentiment ed emotion per spiegare la soddisfazione complessiva nei confronti del docente, misurata come indice di gradimento espresso nei questionari di fine corso che gli studenti sono chiamati a compilare al termine di ogni corso frequentato. Il nostro modello riesce a spiegare dal 49.7% al 72.6% della varianza totale della soddisfazione complessiva verso il docente, inoltre, abbiamo riscontrato che l'emotion joy, quando statisticamente significativo (con $\alpha = 5\%$), aumentana la soddisfazione complessiva verso

Marco Ortu
University Of Cagliari, Dept. of Economics and Business Sciences e-mail: marco.ortu@unica.it

Luca Frigau
University Of Cagliari, Dept. of Economics and Business Sciences e-mail: frigau@unica.it

Giulia Contu
University Of Cagliari, Dept. of Economics and Business Sciences e-mail: giulia.contu@unica.it

l'insegnante, mentre l'emotion anger risulta un indicatore di situazioni con criticità o con ampi spazi di miglioramento.

Key words: Sentiment Analysis, Linear Regression, Natural Language processing

1 Introduction

The spread of Coronavirus disease that shattered the world in 2019 profoundly changed our daily life habits. Universities around the world have had to quickly adapt to a remote mode of teaching and exams. Figure 1 shows the sentiment and emotion extracted by students reports during the last two years, both sentiment and emotion shifted toward negative values, sharpening negative emotions and decreasing positive ones.

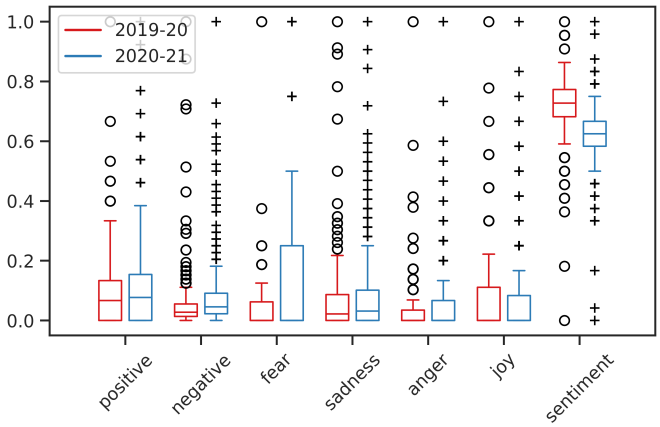


Fig. 1: Emotion and Sentiment for the two years analysed.

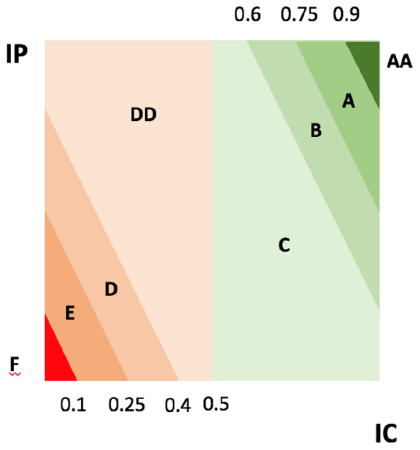
This motivated us to investigate whether the sentiment and emotion expressed by students' reports in the last two years explains the overall satisfaction of the students with their teachers as reported in the end-of-course questionnaires. We could explain from 49.7% to 72.6% of the total variance of three indicators of the overall satisfaction with the teacher, and we found that *joy* emotions, when statistically significant (with $\alpha = 5\%$), show a positive influence on the overall satisfaction with the teacher, while *anger* is a warning signal for critical/improvable situations.

2 Methodology

At the end of each university course, students are asked to fill in a QA questionnaire with ordinal answers: i) definitely NO ii) more NO than YES iii) more YES than NO iv) definitely YES. Questions are grouped in three different areas of interest: i) course subject ii) teaching iii) interest and satisfaction. In the last category we can find the two question considered in this study: i) the overall satisfaction with the subject of the course ii) the overall satisfaction of the student with the teaching. The ordinal answers frequencies are used to evaluate the following QA indexes IC and IP in Equation 1 and 2.

$$IC = \frac{\#YES + \#more\ YES\ than\ NO}{\#totalanswers} \quad (1)$$

$$IP = \begin{cases} \frac{\#YES}{\#YES+\#more\ YES\ than\ NO} & if\ IC > 0.5 \\ \frac{\#more\ NO\ than\ YES}{\#NO+\#more\ NO\ than\ YES} & if\ IC \leqslant 0.5 \end{cases} \quad (2)$$



(a) IC an IP indicators.

(b) Classes Evaluation.

Fig. 2: Student Assessments Evaluation.

The IC an IP indicators are used to identify a point in the IC-IP Cartesian plane, this plan identifies regions to use to define categorical quality indicators showed in Figure 2b. Students may also report issues directly to the QA office of the university, this kind of issues consists in a textual description of the issue associated directly to the teacher in charge of the course.

We combined these two sources of information, linking the textual issues to the overall score of the teacher. To evaluate sentiment and emotion we used two tools for sentiment, one specifically for Italian language and one for English language, and two tools for emotion, one specifically trained for Italian language and one for English language. This multiple language approach is typical for non-English text [1, 5, 6, 3]. For sentiment and emotion in Italian we used the tool provided by Bianchi et al. [2], while for sentiment in English we used VADER tool provided by Hutto et al. [4], and for emotion in English we used the *text2emotion* Python

library[1]. Before the extraction sentiment and emotion in English, we translated the Italian text using the *googletrans* library [2]. We extracted the sentiment and emotion content from the textual description of issues reported by student considering the following sentiment and emotion indicators: # sentence with positive sentiment, # sentence with negative sentiment, # sentence with fear emotion, # sentence with anger emotion, # sentence with sadness emotion and # sentence with joy emotion. In the remainder of this study we will use the *_ENG* suffix for sentiment and emotion extracted using the English tools, while no suffix is used for those extracted using the Italian tools.

We used a multi-linear regression model to explain students' assessments with sentiment and emotions. In particular we selected the three indicators: *IC* index, *DD* and *C*, as response variables of our model, and the sentiment and emotions indicators as predictors. All analysis are performed using the *statsmodels* Python Library[3].We considered a total of 1389 issues reported by students from 2019 to 2021 regarding to 579 teacher and a total of 1169 end-of-course questionnaires.

3 results

We considered three response variables: *IC*, *DD* and *C*; and seven predictors: *positive sentiment*, *negative sentiment*, *joy* (ENG), *sadness* (ENG), *anger* (ENG), *fear* (ENG) and the number of words.

(a) *IC*

R-squared: 0.726		
	coef	**P>|t|**
Intercept	0.8538	0.0000
positive	0.3043	0.0000
negative	0.2725	0.0000
Joy_eng	0.0565	0.5660
Sad_eng	-0.0883	0.0720
Angry_eng	0.3015	0.6550
Fear_eng	-0.0057	0.9240
words	-0.0037	0.0000

(b) *DD*

R-squared: 0.497		
	coef	**P>|t|**
Intercept	-1.5095	0.0010
positive	0.4472	0.0470
negative	1.6246	0.0000
Joy_eng	-5.7262	0.0000
Sad_eng	3.7527	0.0000
Angry_eng	14.8277	0.0130
Fear_eng	-0.3919	0.4580
words	-0.0262	0.0000

(c) *C*

R-squared: 0.507		
	coef	**P>|t|**
Intercept	0.8176	0.1920
positive	1.8982	0.0000
negative	2.5537	0.0000
Joy_eng	0.0635	0.9580
Sad_eng	0.4101	0.4950
Angry_eng	21.6643	0.0090
Fear_eng	-1.1097	0.1300
words	-0.0402	0.0000

Table 1: Multiple Regression Model.

Table 1 shows the linear regression results. We choose the predictors considering a trade-off between R^2 of the model, that is the percentage of explained variance

[1] https://pypi.org/project/text2emotion/

[2] https://pypi.org/project/googletrans/

[3] https://www.statsmodels.org/stable/index.html

by the model, and the interpretability of the model. In general we can see that the number of words reported by students has a negative influence in all the response variables, meaning that the longer the comment the lower the over all satisfaction. The *DD* and *C* represents respectively: slightly critical and just positive situation with great margin for improvement, we can see that these two indicators differs for the *joy* (ENG) indicator which has a complementary effect, positive for *C* and negative for *DD*. Both indicators are driven by anger which presents the higher coefficient, meaning that decreasing the anger results in a shift toward better quality of assessment. The effect of the sentiment is difficult to interpreter as both negative and positive sentiment have a positive influence on the response variables.

(a) *Positive Sentiment*

accuracy 0.76	
Term	**High:Low**
ottima	9.0 : 1.0
confusione	8.4 : 1.0
coordinamento	8.4 : 1.0
puntuale	8.4 : 1.0
complesso	7.8 : 1.0
facile	7.2 : 1.0
falso	6.8 : 1.0
esercizio	6.5 : 1.0
idea	6.5 : 1.0
numerosi	6.5 : 1.0

(b) *Negative Sentiment*

Accuracy 0.63	
Term	**High:Low**
confusione	9.8 : 1.0
dati	8.4 : 1.0
nuovo	8.4 : 1.0
esercizio	7.7 : 1.0
seconda	7.7 : 1.0
visti	7.7 : 1.0
obiettivi	7.0 : 1.0
ordine	7.0 : 1.0
problematiche	7.0 : 1.0
vere	7.0 : 1.0

(c) *Joy*

Accuracy 0.57	
Term	**High:Low**
laboratoriale	8.8 : 1.0
pieno	8.8 : 1.0
certe	7.0 : 1.0
interattiva	7.0 : 1.0
miglior	7.0 : 1.0
autonomo	6.0 : 1.0
formativa	6.0 : 1.0
incoerente	6.0 : 1.0
insoddisfacente	6.0 : 1.0
ottimale	6.0 : 1.0

(d) *Sadness*

Accuracy 0.62	
Term	**High:Low**
passate	13.5 : 1.0
annuale	10.0 : 1.0
logorante	10.0 : 1.0
pubblica	8.0 : 1.0
esaustive	6.5 : 1.0
pieno	6.5 : 1.0
caotica	6.5 : 1.0
informatici	6.5 : 1.0
proibitive	6.5 : 1.0
pochissimo	5.6 : 1.0

(e) *# Anger*

Accuracy 0.64	
Term	**High:Low**
compositiva	6.3 : 1.0
impossibili	6.3 : 1.0
irreperibili	6.3 : 1.0
ostico	6.3 : 1.0
validissima	6.3 : 1.0
accessibile	5.6 : 1.0
accurata	5.6 : 1.0
agonistico	5.6 : 1.0
bassissima	5.6 : 1.0
catastrofico	5.6 : 1.0

(f) *# Fear*

Accuracy 0.65	
Term	**High:Low**
compositiva	9.0 : 1.0
validissima	9.0 : 1.0
accurata	7.8 : 1.0
agonistico	7.8 : 1.0
autoironico	7.8 : 1.0
coeso	7.8 : 1.0
concludenti	7.8 : 1.0
considerevole	7.8 : 1.0
deleterio	7.8 : 1.0
disagevole	7.8 : 1.0

Table 2: Sentiment And Emotion Most Informative Terms.

We further investigated how the tools extract sentiment and emotions. In order to better understand how the sentiment and emotion were evaluated by the different tools, we first classify each text in two categories per sentiment/emotion: text containing high and low level of each sentiment feature. We considered the median of each feature and classified *High*, all text with each feature higher than the median,

and *Low* all texts with lower or equal to the median values. We extracted the top ten most informative terms for *Positive/Negative* sentiment, *joy*, *Sadness* (ENG), *Anger* (ENG) and *Fear* (ENG). In particular we used a Naive-Bayes Classifier to classify the text into two binary categories, *High*/*Low* content of each sentiment and emotion feature. Table 2 reports the results of this classification. In the first column is reported the term contained in a text, and in the second column is reported the odds *High*:*Low* for that term, e.g. for positive sentiment, if a text contains the term "ottima" it is 9 times more likely to contain *High* positive sentiment. For emotions the terms tends to be more coherent, while for negative and positive sentiment we can see some overlap of terms which could explain the result reported in Table 1 for the IC indicator. Furthermore, we can also see how the issues reported by students are extremely heterogeneously and subjective, some peculiar words such as "informatici" are related to a few or single students which perceived some negative aspects.

4 Conclusions

In this study we investigate whether the sentiment and emotion expressed in students' issues reported in the last two years of COVID-19 pandemic, are able to explains the overall satisfaction of the students with their teachers, as reported in the end-of-course questionnaires. We could explain from 49.7% to 72.6% of the total variance of five indicators of the overall satisfaction with the teacher, and we found that positive/negative sentiment emotions, when statistically significant (with $\alpha = 5\%$), show a positive/negative influence on the overall satisfaction with the teacher.

References

1. Balahur, A., Turchi, M.: Improving sentiment analysis in twitter using multilingual machine translated data. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pp. 49–55 (2013)
2. Bianchi, F., Nozza, D., Hovy, D.: Feel-it: Emotion and sentiment classification for the italian language. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 76–83 (2021)
3. Di Rosa, E., Durante, A.: App2check: a machine learning-based system for sentiment analysis of app reviews in italian language. In: SIDEWAYS@ LREC, pp. 8–13 (2016)
4. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8 (2014)
5. Kastrati, Z., Imran, A.S., Kurti, A.: Weakly supervised framework for aspect-based sentiment analysis on students' reviews of moocs. IEEE Access **8**, 106,799–106,810 (2020)
6. Mac Kim, S., Calvo, R.A.: Sentiment analysis in student experiences of learning. In: EDM, pp. 111–120. Citeseer (2010)

Session of solicited contributes SS9 – *Innovation and Value Co-creation in Society*
Organizers and Chairs: Alessandra De Chiara & Anna D'Auria

The smart working towards a Society 5.0

Lo smart working verso una Società 5.0

Sofia Mauro

Abstract The Society 5.0 marks the beginning of a new phase in human history, based on the values of innovation, sustainability and inclusiveness, through the use of technology and forms of cooperation between all of the society's stakeholders. It places the well-being of people at the center of its model, leading to a complete transformation of our lifestyle. The transition to this society is partly due to the spread of the Covid-19 pandemic that hit the whole world in 2020. This aspect was analysed with the reference to the world of work, in which we try to identify the effects of changes and the needs for sustainability, which the Society 5.0 has introduced in our community, starting from the key indicators that include both movements.

Abstract La società 5.0 segna l'inizio di una nuova fase della storia umana, basata sui valori dell'innovazione, della sostenibilità e dell'inclusività, attraverso l'utilizzo della tecnologia e forme di cooperazione tra tutti gli stakeholder della società. Essa pone al centro del proprio modello il benessere delle persone, comportando una completa trasformazione del nostro stile di vita. Il passaggio a questa società, si deve in parte al diffondersi della pandemia di Covid-19 che nel 2020 ha colpito il mondo intero.

[1] PhD in International Studies, University "L'Orientale" of Naples, Italy, email: sofiamauro10@gmail.com

Aspetto analizzato soprattutto dal punto di vista lavorativo, in cui si cerca di individuare gli effetti dei cambiamenti e le esigenze di sostenibilità, che la società 5.0 ha introdotto nella nostra comunità, partendo dagli indicatori chiave che includono entrambi i movimenti.

Key words: Society 5.0, Covid-19, smart working, sustainability

1 Introduction

In recent years, due to globalization and the increase in the level of technical competence of the population, the digital age has experienced significant developments and advances in information technologies. All of this is bringing about rapid changes in the whole vision of the way we live, interact and work within a society. In fact, the development of technologies, the progress of science and the strong independence from digital platforms are leading to the construction of an ideology based on a Society 5.0 and the concept of the Smart City. These aspects will undoubtedly have important consequences also on the working methods adopted so far, particularly favoring the adoption of smart working, compared to a traditional form of work.
The Society 5.0 is in fact by definition a "super intelligent society", in which at the base there is a continuous progress of information and communication technologies, and digital technologies of all kinds, which aim to provide all important opportunities for innovation, growth and prosperity to individuals and society. Obviously, these opportunities will be provided through forms of cooperation and services between man and machine. A hint of this new form of society a work level is due to the Covid-19 pandemic, which has forced many companies to resort to smart working, so that the world economic system does not stop working.
The decision to tackle this issue is given by the objective of actually analyzing whether in a society inclined to innovation and digitization, workers are truly satisfied, both organizationally and emotionally, with the technologies adopted in their sectors and whether their well-being is completely placed at the center of this model.

2 Literature review

2.1 The Society 5.0

The Japanese government in 2016 first brought out the concept of "intelligent society", defined in Society 5.0. The purpose of this society becomes to satisfy the

needs and requirements of all members of a society, by offering goods and services in the required quantity, so that all citizens can live a comfortable life through the provision of high quality services [7]. The Society 5.0 refers to the latest generation, totally digital, in which the tendency to always be connected emerges, providing a system based on three pillars, sustainable and inclusive socio-economic, where the principle of equity is in force, supported by digital technologies, from Big Data Analysis and artificial intelligence [5]. In fact, as Keidanren argues [4], the Society 5.0 can be understood as the "Society of Imagination", in which the digital process can introduce a sustainable society through different values and strong creativity. The ICT technologies can play a fundamental role in empowering citizens, reducing traffic congestion, protecting the environment and combating climate change, effectively responding to the needs of the most disadvantaged communities. In a study presented by Chakravorti and Chaturvedi (2017) [4], the roles of Citizens were also taken into consideration, which must include aspects such as inclusiveness, human condition and development of talent, elements that can overcome prejudices and social barriers. The Society 5.0 could then contribute significantly, as argued (Nakanishi, 2019), to the achievement of the Sustainable Development Goals (SDGs, 2018) adopted by the United Nations.

2.2 The smart working

Twenty-five years ago, the American Jack Nilles coined the term "teleworking" for the first time, on the occasion of the first oil shock. It is therefore not surprising that the initial interest in this issue was driven by concerns related to traffic congestion and increased pollution in densely populated areas [1]. In this definition, the aspect that can be found is the elimination of travel, which is no longer strictly necessary, because it is possible to work remotely in a practical and safe way, thanks to information technologies. In fact, over the years a greater awareness of the opportunities offered by technological progress has emerged, as it is capable of offering immediate access to any information and canceling space-time constraints [8]. Technological progress therefore allows man to acquire new skills, changing his way of working and aiming for innovation based on a more autonomous, flexible and decentralized form of management [3]. In this way, the work has a good chance of being more creative and creative, also achieving sustainability objectives not only environmental, but also economic and social, therefore for the benefit of an entire community.

3 Aim and methodology

The research was designed to investigate the strengths and weaknesses emerging from the adoption of smart working on the worker, both at the organizational level and in his private sphere. At the same time, it intends to investigate what future can be reserved for smart working after the pandemic, for its aspect aimed at guaranteeing

a form of lasting sustainability over time, thanks to the support of digitization, flexibility and advanced communication.
The research data were obtained through interviews with 75 smart-workers, based on semi-structured questionnaires, in which the interviewees shared their experience. The main criterion used in the choice of the interviewees was to identify workers, who had never carried out their work remotely before, if not starting from the first lockdown that took place in Italy in March 2020, in which there is it was a forced total closure.
This study contributed significantly to the research, as the smart-workers interviewed had a yardstick, which allowed them to compare and contrast their previous traditional activity with the remote one, highlighting important considerations. The study was conducted in February and March 2021 in a completely anonymous way and to maintain a form of confidentiality, the names of the organizations and respondents will never be mentioned.

4 Results

4.1 Organizational aspects

At the organizational level (Figure 1), 55% of the interviewees claim to have noticed an improvement, above all for the flexibility of places and times, that this modality entails. Many workers in this period have in fact had the opportunity to get close to their loved ones, leaving the big cities where they previously carried out their work. Some of them have set up their office outdoors, enjoying the sun on fine days, a further positive factor that can foster creativity, unlike the static that can involve working indoors. In fact, there has been an increase in productivity and results, as people have shown maximum commitment since this modality has allowed them to continue working and at the same time stay closer to their families.

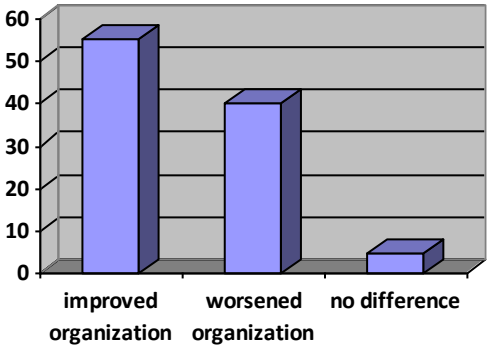


Figure 1: Differences between on-site and remote activities

4.2 Emotional and physical aspects

On an emotional level (Figure 2), the research found that 64% of respondents experienced both positive and negative effects. The most positive aspect that emerged is the reduction of stress, because this mode allows you to work comfortably from home, avoiding the stress of any travel to reach the workplace. However, among the main reasons for discontent we find isolation and the lack of socialization. In fact, half of the interviewees say they have suffered from the lack of opportunities for social contact with colleagues, having the perception of living only and exclusively to work. Equally important is the neglect of oneself and the bad eating habits that have emerged, because the idea of staying at home and being able to take advantage of its comforts has revolutionized the lifestyle of workers.

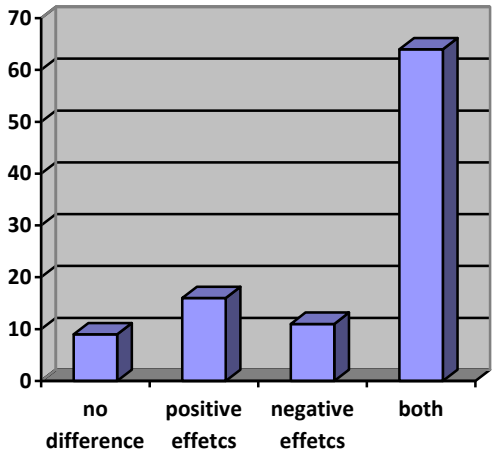


Figure 2: Emotional aspects

From a purely physical point of view (Figure 3), having a sedentary lifestyle, always assuming the same position above all, increases the risk of health problems, which at first the worker did not manifest. In fact, almost all of the interviewees revealed that they suffer from musculoskeletal disorders, headache and visual fatigue.

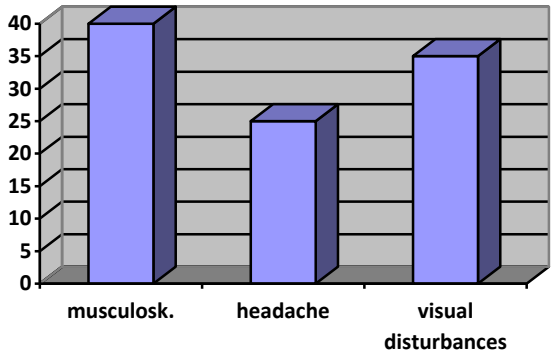


Figure 3: Physical aspects

5 Conclusion

This research unexpectedly revealed a good reaction from the workers interviewed in adopting this new method. Some workers would in fact be able to do everything to adopt this method, others on the contrary, have serious doubts, but applying the right strategy could be a winning solution, such as alternating periods of work on site and others at home, in order to continue to benefit. Positive aspects emerged and counterbalance the negative ones, such as the feeling of strong isolation. All this makes it possible that in the future a mixed working system can be applied based on a type of management that does not aim at the rigidity of the hours, but only at results, in order to guarantee a greater level of well-being.

References

1. Bailey, D.E., Kurland, N.B.: Telework: The Adavantages and Challenges of Working Here, There, Anywhere and Anytime in Accademia.edu, 53-54 (1999).
2. De Chiara, A.: Rapporti professionali e paradigmi dell'impresa digitale. Giappichelli Editore, Torino (2005).
3. Di Nicola, P.: Il nuovo manuale del telelavoro. Seam, Roma (1999).
4. Guevara, A., Rizzi, R.: A ranking of countries concerning progress towards a Society 5.0 in RISUS-Journal on Innovation and Sustainability **11**, 4, 190-192 (2020).
5. Iqbal, A., Olariu, S.: A Survey of Enabling Technologies for Smart Communities in MDPI Journals **4**, 56-59 (2020).
6. Nair, M., Tyagi. A., Sreenath, N.: The Future with industry 4.0 at the Core of Society 5.0: Open Issues, Future Opportunities and Challenges in IEEE Comput. (2021).
7. Shiroishi, Y ., Uchiyama, K ., Suzuki, N.: Society 5.0: For Human Security and Well-Being in IEEE Comput. **51**, 7, 91-92 (2018).
8. Zucaro, R.: Lo smart working: strumento per la conciliazione vita-lavoro e la produttività, Quaderni di pedagogia del lavoro, Pensa, 144-145 (2016).

Innovative interaction in Society 5.0: insight from the cultural sector

Interazioni innovative nella Society 5.0: approfondimenti dal settore culturale.

Erica Del Vacchio, Francesco Carignani, Cesare Laddaga and Francesco Bifulco

Abstract Starting from the concept of Society 5.0, this study aims to analyse how cultural institutions have reacted to Covid-19 and the role of new technologies in the development of resilience dynamics. The data was collected through an online survey to the managers of Italian museums and archaeological parks, highlighting the need to accelerate the digitization processes and to rethink the dynamics related to the digital cultural offer.

Abstract *Partendo dal concetto di Society 5.0, questo studio vuole analizzare come le istituzioni culturali hanno reagito al Covid-19 ed il ruolo delle nuove tecnologie nello sviluppo di dinamiche di resilienza. I dati sono stati raccolti attraverso un sondaggio online a musei e parchi archeologici italiani, evidenziando la necessità di accelerare i processi di digitalizzazione e di ripensare alle dinamiche legate all'offerta culturale digitale.*

Key words: visitors, cyberworld, society 5.0, value creation, covid-19

1 Introduction and Overview

[1] Erica Del Vacchio, University of Naples Federico II, erica.delvacchio@unina.it:
Francesco Carignani, University of Naples Federico II, francesco.carignani@unina.it:
Cesare Laddaga, University of Naples Federico II, cesaere.laddaga@unina.it:
Francesco Bifulco, University of Naples Federico II, francesco.bifulco@unina.it:

Society 5.0 is the fifth phase of the human-centered society already theorized by Moore (1993) with the concept of innovation ecosystem. (Fukuda, 2020).

The technological element is fundamental in the creation of society 5.0 as highlighted in the Japanese case where the national strategy for implementing artificial intelligence has played a crucial role for the country. (Strategic Council for AI Technology, 2017.).

Then, with the publication "Toward realization of the new economy and society - Reform of the economy and society", Keidanren (2016) lays the foundations for the Society 5.0. The scholar defines this new concept as "a human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space" (Keidanren, op. cit., p. 5). Therefore, we are witnessing a transition from the current 4.0 society - in which knowledge and information are not officially shared, with a consequent limitation for the benefits of cooperation in society (Lee et al., 2015) - to the Society 5.0, characterized by a massive use of technologies used to share information and create and co-create value in society (Shiroishi et al., 2018). Other scholars consider Society 5.0 in relation to Industry 4.0, stating that the main purpose of Society 5.0 is to enhance people's quality life with the use of potentialities acquired by Industry 4.0 (Pereira et al., 2020, p.3305).

de Hoyos Guevara et al. (2020) argue that Society 5.0 will be a "Society of the imagination" characterized by a combination of digitization and creativity of different people in order to solve complex problems and create value. This new vision of society provides the possibility to create a Smart Community, a sustainable and human-centered society through the use of new technologies (Iqbal, A., Olariu, S., 2021). The connection between people and things and between real and cyber worlds, which characterizes the new vision of society we are talking about, will allow the effective and efficient resolution of social and complex problems with the possibility of creating a greater quality of life for people and to support healthy economic growth (Pereira et al., 2020). Indeed, the current sanitary emergengy highlight the central role of new technologies and digital tools in ensuring that businesses continue to provide citizen services in compliance with social distancing and anti-Covid regulations (e. g. robots used in healthcare (Khan et al., 2020), mobile health apps and digital platforms (Alexopoulos et al., 2020) the blockchain to track infections (Marbouh et al., 2020).

To discuss the impact of new technologies in the society, with special reference to the current crisis, we adopt the cultural heritage as a research context. Indeed, due to the lockdown, cultural firms and institutions had to re-invent their offer. Studies on the impact of new technologies on the society during the pandemic are still scarce. For hence, our study tries to fill this gap, analyzing how Italian museums and archaeological parks have dealt with this situation of emergency and what role technology has played..

2 Aim and methodology

We identify two research questions: (RQ1) how Italian cultural institutions and museums, as well as archaeological parks have faced the pandemic and if they have managed to be resilient; (RQ2) have new technologies represented a lever for cultural firms and institutions to be resilient. To answer these two research questions, we conducted an analysis based on a case theory method (Gummesson, 2017). The methodology was chosen because it helps the researcher to understand what the individual does in practice and the sociocultural contexts in which he lives (Gummesson, op.cit.)

The data collection was performed through an online survey (30 respondents) to Italian museums and archaeological parks managers. The sample was chosen based on the cultural institutions that were present in the MiC (ex MiBACT) database dedicated to the 'Culture does not stop' initiative, although we focus our attention to those that started a path of digital improvement of communication tools. This sample is divided into three clusters based on geographic location (Northern Italy 19%, Central Italy 33%, Southern Italy and Islands 48%).

3 Findings

As concerns the first RQ, Italian museums and archaeological parks have managed to pursue their strategic choices and deliver their core services through the more massive use of new technologies. and, in particular, of digital communication channels. In fact, 90% of our sample report that their online activity increased during the lockdown period. In particular, the number of contents published monthly has increased: Facebook (increased by 100%), Instagram (increased by 131%), Twitter (increased by 100%). Cultural institutions have tried to be close to citizens through effective and targeted communications, using a multi-channel strategy. In addition, 83% of respondents articulated a heterogeneous offer of ad hoc digital content, proposing new initiatives or, sometimes, strengthening the existing ones, to cope with forced closure and keep the relationship with community alive.

The aforementioned online services had already been implemented in the period prior to the lockdown in 40% of cases, while for the remaining 60% the health emergency represented an impulse to increase their offer of online services, both during and following the lockdown. The forced closure has stimulated creative responses from museums and archaeological parks. Indeed, our sample presented a wide range of ad hoc digital projects and activities, which have never been proposed before, to continue to support access to cultural heritage, keep the relationship with their audiences alive and attract new. It is clear that the in-depth analysis of museum collections and exhibits were the most chosen contents for social media, proposed by 50% of the responding institutions, in the form of posts, photographs and videos.

The Archaeological Park of Paestum and Velia, for example, has published on its Facebook page the "Tales of archeology", daily bulletins edited by the then director, through which to narrate the wonders of ancient Poseidonia, ity of the Magna Greece today Unesco heritage, with footage from the excavations, the museum, the offices and the deposits. The Archaeological Museum of Venice, on the other hand, has implemented a strategy based on information and entertainment: with the hashtags #sentichiparla and #cronachedalmuseochiuso it has given the floor to the works kept in the museum, which tell their own story while the museum is closed to the public. In the Instagram and Facebook stories, on the other hand, every weekend, the #archeomusic column was created, in which a photo of a work, with a caption, associated with a modern music song to explain the meaning in a playful way. Secondly, with an adoption of 35%, there are virtual tours, which are divided into different ways. Often it was video content that showed directors, staff members or museum collaborators wandering around necessarily deserted buildings or archaeological parks and guiding virtual visitors. In this regard, an interesting example is represented by the Egyptian Museum of Turin, which has published on its YouTube channel "The Director's Walks", a cycle of videos that has met with great success with the public, in which you can see the director, which virtually accompanies the spectators in the rooms, illustrating them the most significant objects of the collection and telling their story. Finally, the Archaeological Park of Pompeii has made a virtual tour available on the MiC (ex MiBACT) YouTube channel thanks to the use of a drone that, flying over the ancient city, destroyed by the eruption del Vesuvius in 79 Widespread was the online translation of scheduled events (10%), especially conferences, often offered in the form of live videos on social platforms or websites, the video recordings of which remain available even after some time.

A special reference is due to the National Archaeological Museum of Naples (MANN) which, during the lockdown, transferred the historic review "Meetings of Archeology" to the institution's Facebook page, thus enriching the already multifaceted cultural offer. The MANN, in fact, stood out for its online activity, so much so that it was recognized as the most active Italian museum on Facebook in terms of posts published. The Archaeological Park of Ostia Antica has also transformed the series of events entitled "Let's meet in Ostia Antica", dedicated to public archeology and legality, into videoconferences, accessible from the website.

Furthermore, some museums and archaeological parks have launched initiatives that required and required a higher level of participation by users, such as contests (5%). The public response to the many initiatives put in place by Italian museums and archaeological parks was overall positive. The analysis on the involvement of online audiences showed, in fact, that for 50% of the respondents, the growth rate of the fanbase was higher in the year 2020 than the average of previous years, in relation to the social platform on which the most concentrated online activity. The remaining 50% of the sample is not homogeneous but is divided into two sections: 30% of the interviewees did not actually detect a higher fanbase growth rate than in the past;

20%, on the other hand, said they did not have adequate tools and / or skills to be able to answer the question and were unable to calculate any fanbase growth. Therefore, during the months of blocking of travel, half of the museums considered, in the face of an increase in their publications, found an actual increase in the interest of online audiences, resulting in a significant growth in the number of followers on social pages. Even the engagement rate, in the face of an increase in online activity by museums, showed an increase in the lockdown period compared to the previous period, as was stated by 93% of respondents. Even in this case, the remaining 7% did not deny that there was an increase in the engagement rate but declared it impossible to measure it. This growing trend demonstrates the effectiveness of digital tools in bringing the public closer to the national archaeological heritage.

Between the lockdown period and the immediately following quarter, or between phase 1 and phase 2 of the management of the epidemic, the activity of museums and archaeological parks on social media was not homogeneous between the different institutions and, in parallel, the engagement rate also presented a variable trend. The correspondence between online activity and the performance of the engagement rate.

The health emergency broke out at a time when several institutions had already begun to question innovative ways of offering a cultural experience. 79% of respondents, in fact, declared that they felt the urgency of carrying out long-term planning that included digital innovation, even before the crisis occurred, while 21% understood the need in the face of closure. forced due to the epidemic. However, only 16% of institutions have drafted or plan to draft a strategic plan, at least three years, that includes the development of paid digital cultural offerings. Within this group, the type of non-free cultural offers that have been planned mainly concern the digitization of collections, virtual visits and exhibitions, online learning, apps, and video games.

4 Implications and conclusion

The results of our research show that technology has allowed museum businesses to be resilient and overcome moments of crisis such as the pandemic, as suggested by the scholars (Houston et al., 2021).

This demonstrates how digital can strongly contribute to achieving one of the specific audience development objectives of cultural marketing, that is, also reaching audiences characterized by strong barriers to access, with a view to democratizing cultural use (Antonello et al., 2020). As recent studies (Iqbal, op. cit.) suggest, our results empirically show that we are witnessing the development of a human-centered, sustainable society capable of being resilient and, therefore, preparing for and resist disruptions caused by unplanned events.

From a policy point of view, the study shows how museum institutions were still behind in the digitization of a cultural offer. In this sense, Covid has been a useful expedient to accelerate these processes (Agostino et al., 2021). The data presented should be further analyzed on a qualitative level to understand how much there was a strategy behind these processes or whether in most cases museums just transfer the physical offer online, without fully exploiting the potential of digital, especially with respect to those co-creation processes mentioned above (Orlandi, 2020).

In conclusion, the lockdown experience made the need for a digital transformation for museums and archaeological parks more evident and more urgent, and in part accelerated the process, leading to a rethinking of the cultural offer system. The forced closure, despite the great difficulties it entailed, represented an unprecedented opportunity for the experimentation of numerous digital initiatives which, only thanks to real programming, will be able to transform sporadic interventions, into a true innovation of the offering. Digital innovation has proven to be able to strongly contribute to the achievement of the institutional objectives of disseminating and sharing knowledge, opening new avenues especially in terms of audience development, breaking down the barriers to cultural consumption and making museums more inclusive and accessible to all targets of the public. The contemporary debate indicates a transition of the cultural offer of museums in a "phygital" type direction, in which the online and onsite dimensions will no longer be clearly distinct, but integrated into a single overall experience, which will become the heart of their value proposition.

References

1. Agostino, D., Arnaboldi, M., Lema, M.D.: New development: COVID-19 as an accelerator of digital transformation in public service delivery. Public Money & Management, **41**(1), 69-72 (2021)
2. Alexopoulos, A.R., Hudson, J.G., Otenigbagbe, O.: The use of digital applications and COVID-19. Community mental health journal **56**(7), 1202-1203 (2020)
3. Antonello, V.S., Panzenhagen, A.C., Balanzá-Martínez, V., Shansis, F. M.: Virtual meetings and social isolation in COVID-19 times: transposable barriers. Trends in psychiatry and psychotherapy, **42**, 221-222 (2020)
4. de Hoyos Guevara, A J., Terra, D.M., Portes, J.H., da Silva, J.L.A., Magalhães, K.E.: A Ranking of Countries concerning Progress towards a Society 5.0. Journal on Innovation and Sustainability RISUS **11**(4), 188-199 (2020)
5. Fukuda, K.: Science, technology and innovation ecosystem transformation toward society 5.0. International journal of production economics **220**, 107460 (2020)
6. Houston, M.: Facilitating Digital Transformation for Museum Education in Response to COVID-19. January, 12, (2021)
7. Iqbal, A., Olariu, S: A survey of enabling technologies for smart communities. Smart Cities, **4**(1), 54-77 (2021)
8. Japan Business Federation: Toward Realization of the New Economy and Society. Reform of the Economy and Society by the Deepening of “Society 5.0”, Keidanren, Tokyo (2016)
9. Khan, Z.H., Siddique,A., Lee, C.W.: Robotics utilization for healthcare digitization in global COVID-19 management. International journal of environmental research and public health, **17**(11), 3819 (2020)
10. Lee, J., Bagheri, B., Kao, H.A.: A cyber-physical systems architecture for industry 4.0-based manufacturing systems”, Manufacturing Letters, **3,** 18-23 (2015)

11. Marbouh, D., Abbasi, T., Maasmi, F., Omar, I. A., Debe, M. S., Salah, K., Ellahham, S.: Blockchain for COVID-19: review, opportunities, and a trusted tracking system. Arabian Journal for Science and Engineering, **45**, 1-17 (2020)
12. Moore, J.F.: Predators and prey: a new ecology of competition. Harvard business review, **71**(3), 75-86. (1993)
13. Orlandi, S. D.: Museums web strategy at the Covid-19 emergency times. DigitCult-Scientific Journal on Digital Cultures, **5**(1), 57-66 (2020)
14. Pereira, A.G., Lima, T.M., Charrua-Santos, F.: Industry 4.0 and Society 5.0: opportunities and threats. International Journal of Recent Technology and Engineering, **8**(5), 3305-3308 (2020)
15. Shiroishi,Y., Uchiyama, K, Suzuki, N.: Society 5.0: for human security and well-being", Computer, **51**(7), 91-95 (2018)
16. Strategic Council for AI Technology. (2017). Artificial intelligence technology strategy.

Society 5.0: a bibliometric analysis
Society 5.0: un'analisi bibliometrica

Anna D'Auria* and Alessandra De Chiara#

Abstract The present paper aims at discussing the topic of Society 5.0 as it is recently proposed by institutions, practitioners, and scholars as the new configuration of a 'Super Society'. To accomplish this aim, we performed an analysis on the theoretical contributions adopting a bibliometric method with the employment of the software Bibliometrix. The results led us to identify and depict the main pillars of the Society 5.0, namely the pivotal role of citizens, the relevance of new technologies and the sustainable development as the main goal.

Abstract *Il presente articolo si propone di discutere il tema della Società 5.0, recentemente proposto da istituzioni, professionisti e studiosi come la nuova configurazione di una 'Super Società'. Per conseguire tale obiettivo, abbiamo eseguito un'analisi sui contributi teorici adottando un metodo bibliometrico con l'impiego del software Bibliometrix. I risultati ci hanno condotto ad identificare e descrivere i pilastri principali della Società 5.0, ovvero il ruolo cardine dei cittadini, la rilevanza delle nuove tecnologie e lo sviluppo sostenibile come obiettivo principale.*

Key words: Society 5.0, Bibliometric analysis, Citizens, New technologies, Sustainable development

* Anna D'Auria, University "L'Orientale" of Naples, Via Chiatamone, 61/62, 80121, Naples, Italy, adauria@unior.it

Alessandra De Chiara, University "L'Orientale" of Naples, Via Chiatamone, 61/62, 80121, Naples, Italy, adechiara@unior.it

1 Introduction and theoretical overview

The smartization of urban contexts is one of the hottest topics in the international academic debate [1,2,3]. The present study is framed in this research context, with particular reference to the theme of Society 5.0. Society 5.0 is a concept introduced by the Japanese government in 2016 with the aim of defining the society of the present and the near future as "a new type of society where innovation in science and technology occupies a prominent place, with the aim of balancing social and societal issues that need to be solved, while ensuring economic development" [4: 190].

Issues as co-creation of value, dissemination and enhancement of knowledge, innovation, sustainability, and the role of individuals in the society are the fundamental aspects of this new perspective [5]. Indeed, the elements that characterize this configuration are innovation and sustainability as transversal elements to all the societal initiatives, apart from the pivotal role held by people.

Several scholars have set themselves the goal of describing the role of citizens, an emblematic example is that represented by Paskeviciute and Anderson [6] which speak of 'micro behavior' to highlight the relevance of the single individual actions. Still with reference to the role of people, the concept of co-creation can be found in numerous definitions, mainly to underline the network of relationships that is at the basis of the functioning of modern society. Advanced technologies, the global digitization process, the innovative approach that characterizes an increasing number of companies and the same institutions that govern urban contexts and countries, belong to what is called 'Industry 4.0' that has its pillars in the so-called enabling technologies, that create the conditions that allow the configuration of new business models for the improvement of the performance of companies. In addition, as more and more theoretical, and empirical evidence has shown for some time, the new technologies are able to improve the lives of individual citizens and society in general.

In line with this, the Society 5.0 has its main objective in the collective or even global well-being. This approach is closely linked to that of industry 4.0 mentioned above, as also in this case elements such as the digital transition, open innovation and the co-creation of value play a very important role. The link between industry 4.0 and Society 5.0 is highlighted by Aquilani and colleagues, stating that "Industry 4.0 can greatly support the transition to Society 5.0, a society with sustainability at its core, thanks to its features and enabling technologies (i.e., big data, AI and IoT). [5: 2].

2 Aim and research process

Society 5.0: a bibliometric analysis

The novelty of the theme led us to examine academic contributions on the subject, observing what are the aspects most treated by scholars and that emerge as crucial elements of the so-called Society 5.0. To accomplish this aim, we conduct a bibliometric analysis, starting from the Web of Science-Core collection database.

To shape the dataset, we used one single query, namely "Society 5.0", as our aim was to identify the main topics related to the concept. We adopt a bibliometric approach, a "quantitative approach for the description, evaluation, and monitoring of published research" [7: 1]. We used the software Bibliometrix [8], as it allows to perform a broad analysis on the keywords and their linkages and recurrence in all the theoretical contributions. The results have been represented in a graphical form through the software VosViewer [9]. The investigation was carried out with a co-words analysis based on the technique of co-occurrences of terms (co-word analysis) by relating two or more words and considering their co-occurrence in the documents. The first results of the bibliometric analysis are summarized in the following table, "main information".

Table 1: Main information about data from the bibliometric analysis (from Bibliometrix)

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2016:2022
Sources (Journals, Books, etc)	147
Documents	307
Average years from publication	1,21
Average citations per documents	1,218
Average citations per year per doc	0,6088
References	9108
DOCUMENT TYPES	
Article	97
article; early access	10
article; proceedings paper	3
editorial material	3
proceedings paper	185
Review	9

3 Results

As previously anticipated, the results of the bibliometric analysis have depicted three main trajectories of study precisely corresponding to the crucial elements of Society 5.0, namely the role of citizens and their relationships, new technologies, and the sustainable development.

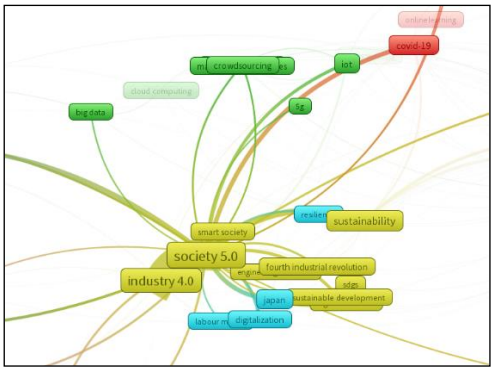


Figure 1: Main results in graphical form (from VosViewer)

3.1 The role of citizens in the Society 5.0

As concerns the role of citizens, they represent the core of the whole society, in line with the 'citizens science' perspective. In fact, Society 5.0 finds in the citizen, or rather in the community, its pivot element [10], highlighting that people are at same time co-creators of services able to improve the quality of life and recipient of the goal to achieve the collective well-being [5,11,12]. In this regard, an interesting contribution is the one provided by Iqbal and Olariu, who claim that "the stated goal of Society 5.0 is to meet the various needs of the members of society through the provisioning of goods and services to those who require them, when they are required and, in the amount, required, thus enabling the citizens to live an active and comfortable life" [13].

3.2 Innovation and new technologies in the Society 5.0

Economic growth and the development of new technologies are interconnected, leading together to the transition to innovative economies that no longer focus only on companies but on the single individual who, thanks to digital devices and new media, take on new skills [14,15]. In the gradual and increasingly rapid development of new technologies, contemporary society overcomes the so-called industry 4.0 reaching the 5.0 phase, considered as the next step in social development based on data-driven innovation [16].

The goal of Society 5.0 is to provide people with advanced services to improve their quality of life thanks to the employment of new technologies. To achieve this goal, it

is necessary to introduce next-generation technologies, reform corporate governance and develop strategies for creating value for cities and communities [5,17].

3.3 Sustainability in the Society 5.0

As proposed by the Japanese government, the Society 5.0 is a society able to face global environmental crisis [18,19,20]. According to the original model, Society 5.0 aims at a constant improvement of technologies, not only digital, in order to prevent, contain, and face the tragic events that can dramatically affect people's lives. Hopefully, this strategy will contribute to the development of a form of sustainable adaptability of society to the external environment, favoring the establishment of a close relationship between the community and the territory [21]. Of course, it should also be highlighted that, although new technologies and this new orientation can favour the achievement of a better and safer world, the so-called enabling technologies can also determine risks and threats to society, for example, the difficulty in protecting privacy, and the possible cybercrimes [22].

4 Discussion and conclusion

When scanning literature contributions, it emerges that most of scholars consider the Society 5.0 as founded on three main pillars, namely the citizens and their participation to the city management and governance in a co-creation perspective, the new technologies, as the lever for the development of the community, and the sustainable development conceived as the main goal of all the actors taking part to the society [10,16,21].

It is important to underline that new technologies and the implementation of innovation processes cannot be considered sufficient for the configuration of initiatives and strategies aimed at the sustainable development, let alone as a proactive response to the serious disasters. In fact, also in line with the models on the smartization of territories that have been proposed in the recent decades [1,2,3], the role of citizens remains fundamental [10], and, above all, the cooperation among a variety of actors that directly or indirectly can affect the development of society.

In conclusion, society 5.0 has as its main objective in responding proactively and resiliently to the constant challenges of contemporary society thanks to the new technologies and an approach that has as its main goal in the sustainable development [18,19,20].

References

1. Giffinger, R., Fertner, C., Kramar, H., Meijers, E.: City-ranking of European medium-sized cities. Cent. Reg. Sci. Vienna UT (2007)
2. Bifulco, F., D'Auria, A., Amitrano, C.C., Tregua, M.: Crossing technology and sustainability in cities' development. Sustainability Science **13**(5),1287-97 (2018)
3. Macke, J., Sarate, J.A., de Atayde Moschen, S.: Smart sustainable cities evaluation and sense of community. Journal of Cleaner production **239**, 118103 (2019).
4. de Hoyos-Guevara, A.J., Terra, D.M., Portes, J.H., da Silva, J.L., Magalhães, K.E.: A Ranking of Countries concerning Progress towards a Society 5.0. Journal on Innovation and Sustainability RISUS **11**(4),188-199 (2020).
5. Aquilani, B., Piccarozzi, M., Abbate, T., Codini, A.: The role of open innovation and value co-creation in the challenging transition from industry 4.0 to society 5.0: Toward a theoretical framework. Sustainability **12**(21), 8943 (2020)
6. Paskeviciute, A., Anderson, C.J.: Macro-politics and micro-behavior: Mainstream politics and the frequency of political discussion in contemporary democracies. In Social logic of politics: personal networs as contexts for political behavior), pp. 228-248. Temple University Press. (2005)
7. Zupic, I., Cater, T.: Bibliometric methods in management and organization: A review. In Academy of Management Proceedings **1**, 13426. Academy of Management. (2013)
8. Aria, M., Cuccurullo, C.: bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of informetrics **11**(4), 959-75 (2017)
9. Van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523-38 (2010)
10. Fukuda, K.: Science, technology and innovation ecosystem transformation toward society 5.0. International journal of production economics **220**, 107460 (2020)
11. Siriweera, A., Naruse, K.: Survey on Cloud Robotics Architecture and Model-Driven Reference Architecture for Decentralized Multicloud Heterogeneous-Robotics Platform. IEEE Access **9**, 40521-39 (2021)
12. Shiroishi, Y., Uchiyama, K., Suzuki, N.: Society 5.0: For human security and well-being. Computer **51**(7), 91-5 (2018)
13. Iqbal, A., Olariu, S.A.: survey of enabling technologies for smart communities. Smart Cities **4**(1), 54-77 (2021).
14. Sawaragi, T., Horiguchi, Y., Hirose, T.: Design of Productive Socio-Technical Systems by Human-System Co-Creation for Super-Smart Society. IFAC-PapersOnLine **53**(2), 10101-8 (2020)
15. Zhanna, M., Nataliia, V.: Development of Engineering Students Competencies Based on Cognitive Technologies in Conditions of Industry 4.0. International Journal of Cognitive Research in Science, Engineering and Education **8** (2020)
16. Fukuda, K.: Science, technology and innovation ecosystem transformation toward society 5.0. International journal of production economics **220**, 107460 (2020)
17. Kitsuregawa, M.: Transformational Role of Big Data in Society 5.0. In 2018 IEEE International Conference on Big Data (Big Data), pp. 3-3. IEEE Computer Society (2018)
18. Goda, Y., Suzuki, K., Kashihara, A.: Technology and Education in Japan: Research, Practice, and More. In2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), pp. 1-3. IEEE (2020)
19. Acioli, C., Scavarda, A., Reis, A.: Applying Industry 4.0 technologies in the COVID–19 sustainable chains. International Journal of Productivity and Performance Management (2021)
20. McLaren, G.: Why the Future Needs Ecological Civilization and Not Society 5.0. Cosmos and History: The Journal of Natural and Social Philosophy **17**(1), 567-98 (2021)
21. Carayannis, E.G., Dezi, L., Gregori, G., Calo, E.: Smart environments and techno-centric and human-centric innovations for Industry and Society 5.0: A Quintuple Helix Innovation System view towards smart, sustainable, and inclusive solutions. Journal of the Knowledge Economy, 1-30 (2021)
22. Iwahana, K., Takemura, T., Cheng, J.C., Ashizawa, N., Umeda, N., Sato, K., Kawakami, R., Shimizu, R., Chinen, Y., Yanai, N.: MADMAX: Browser-Based Malicious Domain Detection Through Extreme Learning Machine. IEEE Access **14(**9), 78293-314 (2021)

Session of solicited contributes SS10 – *Statistical learning for mainstream press, health and fiscal data*
Organizer and Chair: Claudio Conversano

The regression trunk model for partitioning Italian municipalities based on their fiscal capacities and its determinants

Utilizzo del metodo regression trunk per la ripartizione dei Comuni italiani in base alla loro capacità fiscale ed alle sue componenti

Alessio Baldassarre and Danilo Carullo

Abstract The long-term economic performance of a territory is strictly related to the fiscal capacity concept. In Italy, it is calculated for every municipality by the Department of Finance (Ministry of Economy and Finance) to distribute resources based on their expenses needs too. Here, we propose a new way to analyze fiscal capacity and its determinants. Through a particular regression tree model, the so-called regression trunk, it is possible to consider both the main and interaction effects of predictors on determining fiscal capacity. The result is a small regression tree that creates an easy-to-read partition of the Italian municipalities. This model can be helpful when there is no information a priori about what interactions should be added to the model. In addition, we show the effects of each socio-economic predictor on predicting fiscal capacity.

Abstract *Il benessere economico di un territorio è strettamente legato al concetto di capacità fiscale. In Italia la capacità fiscale viene calcolata per ogni singolo comune dal Dipartimento delle Finanze e, assieme ai fabbisogni standard, rappresenta uno dei fattori che vengono considerati dal governo centrale per il trasferimento di risorse ai comuni stessi. In questo articolo proponiamo un nuovo approccio per l'analisi delle determinanti della capacità fiscale. Nello specifico, abbiamo applicato un modello di regressione ad albero, chiamato regression trunk, col fine di ripartire i comuni italiani in base agli effeti principali ed agli effetti interazione tra le variabili prese in considerazione. Il risultato è un piccolo albero di regressione, semplice da interpretare e che può risultare particolarmente utile quando non si hanno informazioni a priori circa le interazioni da considerare nel modello.*

Key words: STIMA algorithm, regression tree, interaction effects

Alessio Baldassarre
Ministero dell'Economia e delle Finanze, Roma, e-mail: alessio.baldassarre@mef.gov.it

Danilo Carullo
Ministero dell'Economia e delle Finanze, Roma, e-mail: danilo.carullo@mef.gov.it

1 Introduction to fiscal capacity

Fiscal capacity can be described as the ability of a country to collect revenues to provide public goods and achieve the typical governments' functions. Generally, fiscal capacity is referred to as tax capacity, even if taxes may be only a part of a government's source of revenue (Kaldor, 1963). Higher values translate in the ability of the government to raise revenues. Usually, developed countries with more vital tax administrations and the ability to enforce tax policies show a higher fiscal capacity (Besley and Persson, 2012; Rogers and Weller, 2013).

Several studies have been conducted on fiscal capacity affecting long-term economic performance and institutional quality (Besley and Persson, 2010; Dincecco and Katz 2016; Dincecco and Prado, 2012). In addition, it represents the resource for implementing social protection programs, ensuring the maintenance of an efficient bureaucracy and the success of policies (Papadia, 2016).

In the Italian political system, fiscal capacity is calculated for every municipality by the Department of Finance. It represents the potential revenue in the reference area. Once fixed a standard tax rate for the municipalities, fiscal capacity is used together with the municipalities' standard needs for distributing a portion of the municipal solidarity fund (Fondo di Solidarietà Comunale). The calculation is conducted by considering two main revenue sources:

1. Real estate taxes (IMU and Tasi), additional municipal income tax on natural persons (Addizionale comunale IRPEF);
2. Minor taxes.

Tax sources in the first category are estimated by following the Representative Tax System (RTS) method since it is possible to calculate their standard level of revenue. The RTS method fits the need of calculating the amount of revenue that a municipality can potentially collect, starting from the relevant tax bases and the legal tax rate. The use of actual or collected revenue requires adjustments to avoid distortions in favor of those municipalities where policies against tax evasion are not effectively pursued. Hence, the tax gap is estimated as the difference between potential theoretical revenue and actual revenue.

Minor taxes' revenues constitute the residual fiscal capacity, and they have to be estimated due to the difficulty of quantifying their tax bases. For this reason, minor tax revenues are estimated through econometrics techniques as the regression-based fiscal capacity approach (RFCA, Di Liddo et al., 2016).

This work presents an innovative application in which the fiscal capacity of Italian municipalities' represents the statistic units for fitting a particular regression tree model. The main goal is to obtain an easy-to-read partition of the Italian municipalities by finding fiscal capacity determinants (main and interaction effects). Specifically, we propose a partition of Italian municipalities by following the regression trunk approach (Dusseldorp and Meulman, 2004) within the STIMA framework (Conversano et al., 2017). This model combines a multiple regression model and a regression tree to discover the interaction effects between variables (over and above their main effects) when there is no information a priori. In addition, results can be

helpful for both descriptive and predictive purposes, representing a valuable tool for policymakers to implement policies aimed at distributing resources from the central government based on the calculation of fiscal capacity.

The sequent sections are organized as follows: the main concepts related to the regression trunk approach are shown in Section 2; then, Section 3 is where we show the main structure of the data set on which we applied the model; finally, in Section 4 we present the results of our analysis in terms of model coefficients output and by reporting the final tree which shows the Italian municipalities' final partition.

2 Regression trunk approach and STIMA algorithm

In a statistical perspective, when the individual effect of two variables do not combine additively, which means they have a joint effect (Cohen et al., 2003), then interaction occurs (Berrington de Gonzàlez and Cox, 2007). The interactive structure between variables can be treated by fitting tree-based models, such as Classification and Regression Trees (CART, Breiman et al., 1984). In the classification community, there are several implementations of tree-based models. Almost all of these models work by treating interactions as threshold interactions, which corresponds to splitting observations concerning the effect of a predictor on a response variable.

Our work treats the fiscal capacity of Italian municipalities as the response variable, and for partitioning our observations, we followed the regression trunk model. It combines a multiple regression model and a regression tree (Dusseldorp and Meulman, 2004) and represents a suitable choice when there are no exact a priori hypotheses about the number and order of interaction effects. Joint effects are estimated over and above their separate effects, and the final result is usually a small regression tree.

The regression trunk model can be formulated as a single linear model.

$$g(\mu) = \eta = \beta_0 + \sum_{p=1}^{P} \beta_p x_p + \sum_{t=1}^{T-1} \beta_{P+t} I\{(x_1, \ldots, x_P) \in t\} \tag{1}$$

The formulation refers to a standard GLM presenting a linear predictor η such that $\mu = g^{-1}(\eta)$ (i.e., μ is an invertible and smooth function of η). The first P parameters represent the main effects part of the model estimated in the root node of the trunk, while the other $T-1$ parameters define the interaction effects part of the model obtained by partitioning recursively in a binary way the observations. The regression trunk model works by adding interaction terms defined by the coefficients β_{P+t} and the indicator variables $I\{(x_1, \ldots, x_P) \in t\}$ in a recursive way. During this process, one node T is considered a reference category for the other interaction effects.

The regression trunk model is implemented within the STIMA framework. A recursive partitioning algorithm creates binary splits to the trunk by adding a new threshold interaction effect. The best splitting variable (and his value) x_p^* is the

one that maximizes the effect size, which is computed as the relative increase in variance-accounted-for. Hence, an additional interaction effect is included if the effect size between the model at the current split and the model including the candidate interaction is maximized. Once the split is found, a dichotomous variable is added to the model, and all regression coefficients are re-estimated. To avoid the overfitting issue, at the end of the regression trunk, the pruning procedure is applied with the cross-validation procedure.

2.1 Application: data and results

In this work, we apply the regression trunk model to administrative data for the year 2018. In the specific, the per capita fiscal capacity of the Italian municipalities represents our analysis's observations. Our dataset is composed of 6629 municipalities belonging to the regions with the ordinary statute. Municipalities subject to territorial variations (mergers, demergers, etc.) were not taken into consideration. Fiscal capacity is calculated as specified in the introduction section. Then, the per capita value comes from considering the population of each municipality for the year 2018. The predictors x_p are chosen from a large set of variables by considering different socio-economic aspects for the year 2018 from the Italian National Institute for Statistics (ISTAT) databases. The chosen variables and their key statistics are shown in Table 1.

Table 1 Variables' key statistics

	vars	NAs	mean	sd	median	min	max	skew	kurtosis	se
Per capita fiscal capacity	y	0	373.88	296.27	323.73	0.00	6965.46	6.64	81.93	3.64
Per capita incoming commuters	x_1	24	0.16	0.15	0.12	0.00	3.80	5.65	87.98	0.00
Illiterate rate	x_2	688	0.85	1.30	0.33	0.00	14.24	2.90	11.03	0.02
Graduates rate	x_3	26	7.13	2.67	6.77	0.00	27.31	1.29	4.32	0.03
Real estate market value	x_4	111	705.93	363.35	640.76	0.00	6480.39	3.97	33.61	4.46
Per capita current expenditure	x_5	86	60.44	17.10	65.31	0.00	93.50	-1.11	1.07	0.21
Per capita total net income	x_6	24	13776.35	13317.18	3407.85	0.00	35909.27	-0.09	0.92	42.11
Digital divide	x_7	24	21.40	32.43	2.20	0.00	100.00	1.37	0.36	0.40

Model output Table 2 shows that all the variables chosen for the analysis significantly affect determining per capita fiscal capacity. Except for the sixth region R_6, all the regions have a significant effect on our response variable. The first region R_1 represents the reference category for interactions. The best interactions in terms of predicting the per capita fiscal capacity for Italian municipalities are: real estate market value (x_4) $\times$ illiterate rate (x_2) $\times$ total net income ($x6$), and real estate market value $\times$ illiterate rate $\times$ total net income $\times$ incoming commuters ($x1$).

Table 2 Regression trunk model: coefficients' output

	Coefficient	Standard Error	Std. Coef.	t value	Pr(>|t|)
Intercept	76.20	19.21	0.00	3.97	0.00
Per capita incoming commuters	75.54	23.21	0.04	3.26	0.00
Illiterate rate	-16.97	3.13	-0.07	-5.43	0.00
Graduates rate	-8.77	1.29	-0.08	-6.78	0.00
Real estate market value	0.17	0.01	0.21	14.15	0.00
Per capita current expenditure	-1.02	0.19	-0.06	-5.38	0.00
Digital Divide	1.05	0.10	0.12	10.45	0.00
Per capita total net income	0.02	0.00	0.22	11.16	0.00
R2	488.96	32.10	0.22	15.23	0.00
R3	140.35	16.55	0.12	8.48	0.00
R4	571.72	27.73	0.23	20.62	0.00
R5	66.80	13.05	0.08	5.12	0.00
R6	-14.43	12.98	-0.02	-1.11	0.27

The regions are defined as follows

$$
\begin{aligned}
&\mathrm{R}_1 = I(x_6 \leq 11198),\\
&R_2 = I(x_6 > 11198, x_2 > 0.009, x_4 > 1687),\\
&R_3 = I(x_6 > 11198, x_2 > 0.009, x_4 \leq 767),\\
&R_4 = I(x_6 > 11198, x_2 > 0.009, x_4 > 767),\\
&R_5 = I(x_6 > 11198, x_2 > 0.009, x_4 \leq 1687),\\
&R_6 = I(x_6 > 11198, x_2 > 0.009, x_4 > 1687),
\end{aligned}
$$

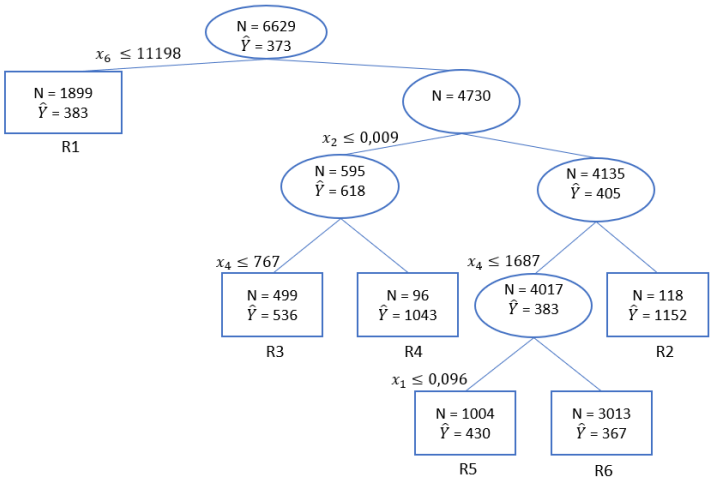


Fig. 1 Final regression trunk: we report for each node of the tree, the number of Italian municipalities N, the estimated fiscal capacity $\hat{Y}$, and the regions R_t for each T terminal node.

By comparing the β-coefficients of illiterate rate and graduates rate, they have a negative effect on fiscal capacity even if the illiterate rate effect is twice the graduates' rate effect. Then, per capita incoming commuters strongly affect the determination of fiscal capacity, while per capita total net income has a low effect on it. The final regression trunk, after the pruning procedure, is shown in Figure 1 and is composed of five splits and six terminal nodes. It represents an easy-to-read interpretation of the results, making the regression tree models widely used to disseminate the results for the public with different backgrounds.

Acknowledgements Opinions expressed in this paper are those of the authors and do not necessarily reflect views of the public administration of affiliation.

References

1. Berrington de Gonzalez, A., and Cox, D. R.: Interpretation of interaction: A review. Ann. Appl. Stat **1 (2)**, 371–385. (2007)
2. Besley, T., and Persson T.: The Origins of State Capacity: Property Rights, Taxation, and Politics. American Economic Review **99 (4)**, 1218–1244. (2009)
3. Besley, J., and Persson, T.: Public Finance and Development. Draft Chapter for the Handbook of Public Economics. (2012)
4. Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J.: Classification and regression trees. Boca Raton, FL: CRC press. (1984)
5. Cohen, J., Cohen, P., West, S. G., and Aiken, L. S.: Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erbaum Associates Inc., Mahwah NJ. (2013)
6. Conversano, C., and Dusseldorp, E.: Modeling threshold interaction effects through the logistic classification trunk. Journal of Classification **34 (3)**, 399–426. (2017)
7. Di Liddo, G., Longobardi, E., Porcelli, F.: Measuring horizontal fiscal imbalance: the case of Italian municipalities. Local Government Studies **42**, 1–35. 10.1080/03003930.2016.1150836. (2016)
8. Dincecco, M., and Katz, G.: State Capacity and Long-Run Economic Performance. The Economic Journal 126. (2012)
9. Dincecco, M, Prado, M.: Warfare, fiscal capacity, and performance: an empirical investigation. (2010)
10. Dusseldorp, E., Conversano, C., Van Os, B. J.: Combining an additive and tree-based regression model simultaneously: Stima. Journal of Computational and Graphical Statistics **19 (3)**, 514–530. (2010)
11. Dusseldorp, E. Meulman, J. J.: The regression trunk approach to discover treatment covariate interaction. Psychometrika **69 (3)**, 355–374. (2004)
12. Herbst, J.: War and the State in Africa. International Security **14 (4)**, 117—139. doi:10.2307/2538753. (1990)
13. Kaldor, N.: Taxation for Economic Development. The Journal of Modern African Studies **1 (1)**, 7—23. doi:10.1017/S0022278X00000689. (1963)
14. Rogers, M., Weller, N. . Income taxation and the validity of state capacity indicators. Journal of Public Policy **34**, 183–206. (2013)

An analysis of Italian healthcare mobility through a depth-based clustering procedure

Un'analisi della mobilità sanitaria Italiana attraverso un metodo di clustering sferico basato sulle data depth

Giuseppe Pandolfo and Carmela Iorio

Abstract This work is aimed at offering a non-parametric approach for the analysis of Italian inter-regional healthcare mobility. Specifically, the proposal exploits the concept of "data depth" for performing cluster analysis of directional data.
Abstract *Lo scopo di questo lavoro è offrire un approccio non parametrico per l'analisi della mobilità sanitaria interregionale italiana. Nello specifico, la proposta si basa sull'utilizzo delle funzioni "data depth" per la cluster analisi di dati direzionali.*

Key words: Data depth, directional statistics, PAM algorithm

1 Introduction

In the recent years the inter-regional mobility has attracted the interest of many researchers, and the interest is still receiving a growing attention. The National Health Service in Italy is a regionally decentralized system, in which patients may choose to receive healthcare services for free at the point of consumption, within a tax-funded system. Even though every region should be able to meet the health needs of its own inhabitants, consistent migration among regions exists. The analysis of patient flows can be useful for health planning purposes, providing precious information about citizens' preferences, and helping health managers to think about inequality and adjust the offer of care.

Giuseppe Pandolfo
Department of Economics and Statistics, University of Naples Federico II, e-mail: giuseppe.pandolfo@unina.it

Carmela Iorio
Department of Economics and Statistics, University of Naples Federico II e-mail: carmela.iorio@unina.it

Our idea is to use data about how hospitalizations of patients of a given region of residence are distributed over the twenty-one Italian regions. Hence, each data vector contains components which are the proportions of some whole. Such data are called compositional data, which are non-negative constant-row-sum data collected in an $I \times J$ matrix where the explanatory variable is characterized by I categories, while the response variable is characterized by J categories. Due to this non-negativity and constant-row-sum, interest usually goes out to, for each row in the matrix, relative quantities for the different column categories that add up to one.

Here we propose to projected compositional data onto the surface of a unit $(d-1)$-dimensional hypersphere and treat them as directional/spherical data. Then, a non-parametric method based on depth functions is used to conduct cluster analysis of such data.

In the remainder of the paper is organized as follows. In Section 2 we offer a brief introduction to directional data (and the connection with compositional data) and data depth. In Section 3, the depth-based procedure is proposed. In Section 4, we analyze the Italian healthcare mobility by means of the proposed method. Finally, some remarks are offered in Section 5.

2 Background

This section is dedicated to offer a brief introduction to directional data and data depth functions.

2.1 Directional data

Directional data are data having an angular nature. Such data arise when observations are directions and are analyzed by means of unit vectors in $\mathbb{R}^d$. As a consequence, data are constrained to lie on the surface of the unit $(d-1)$-dimensional hypersphere $\mathbb{S}^{d-1} := \{x : \|x\|_2 = 1\}$ where $\|x\| := \sqrt{\sum_{i=1}^{d} x^2}$, with $x = (x_1, \ldots, x_d)'$. Statistical tools must take into account that the angular value depends on the choice of the zero direction and on the sense of rotation (statistical procedures must be "rotational invariant"). In addition their occurrence is periodic and no natural ordering of observations exists. Such peculiar features make the use of classical statistical methods inappropriate, and often misleading. In this regard, consider the angles 0 and 2π on a circle. Their arithmetic mean is π, but they are actually the same angle and the "true" mean is 0. Thus, working with such data requires specific techniques that consider the geometry of the manifold, and this holds true also for clustering issues.

2.1.1 Directional data and compositional data

In 1982 Stephens [4] analyzed the connection between compositional and directional data, which is given by the square-root transformation

$$(x_1,\ldots,x_d) \mapsto (\sqrt{x_1},\ldots,\sqrt{x_d}), \tag{1}$$

this way data lie on the positive orthant of the $(d-1)$-dimensional hypersphere. One advantage of this approach over the log-ratio transformation approach is the treatment of zero values components. Hence, this opens the possibility of exploiting tools for directional statistics to model compositional data as well (see, e.g. [3]).

2.2 Data depth concept

A data depth function is aimed at ordering points in a space according to their "centrality" or "depth" with respect to a distribution F (the larger the depth of a point x, the more central x is with respect to F), and it is generally denoted by $D(x,F)$. It is a useful tool in the multidimensional case as well as in the spherical context where ordering the points from the inner to the outer part of a distribution or sample is not a trivial task (see [1] and [2]). Hence, such functions offer a chance to exploit non-parametric methods in multivariate data analysis. Many depth measures with different characteristics can be found in the literature, and the field of applications of data depth is vast and still growing. For the purpose of this work we adopt the notion of cosine distance depth which has been introduced by [2] to analyze directional data.

3 Proposal

The aim is thus finding clusters to describe the mobility trends among the Italian regions in the case of ordinary hospitalizations. In this regard, the square root transformation defined in (1) allows us to analyze data as directional data through a data depth function for directional data. Specifically, a depth-based medoid algorithm is adopted by using the notion of cosine distance depth (CDD) on the basis of two following main concepts.

- ***Depth based partition:*** a depth-partition C_k is a non-empty subset based on the depth values of data set X of dimension $n(d-1)$, defined in $\mathbb{S}^{d-1}$, such that $X = c_1,\ldots,C_k,\ldots,C_K$, where $K \leq n$.
- ***Depth medoid:*** a depth medoid $X_{D_k} \in \mathbb{S}^{d-1}$ is a point belonging to a depth-partition C_k such that:

$$x_{D_k} = \arg\max_{x \in C_k} D(x,X).$$

In other words, the depth medoid is the point with the highest depth value within the k-th depth partition.

The procedure follows the partition around medoids (PAM) algorithm. The marked difference is that we iteratively search for those points (the depth medoids) which maximize a given depth function between the depth medoids themselves and other points belonging to the same partition. Let X be a given data set of dimensions $n \times (d-1)$ defined in $\mathbb{S}^{d-1}$. We randomly select K observations as depth medoids and perform the following steps iteratively:

1. Assign each point to the cluster for which the depth function between the point and the K depth medoids is maximum;
2. Redefine the K depth medoids by selecting those points that have the highest depth within the cluster.

Such two steps are repeated until there is no longer a new assignment to the clusters.

4 Analysis of the Italian healthcare mobility

Data refer to hospital admissions in Italy as reported by the Italian Ministry of Health annual reports (https://www.salute.gov.it/portale/home.html) in 2017. Such data represent the tool for collecting information relating to all hospitalization services provided in accredited public and private hospitals present throughout the national territory. We collected data for all Italian regions with attention to the mobility of patients among regions. This because the analysis of inter-regional healthcare mobility represents one of the main criteria for the evaluation of the Regional Healthcare Systems, both in terms of its economic-financial relevance and the quality and satisfaction of the services provided. Here we focus on acute care admissions for the twenty-one Italian regions.

The goal is to identify a structure of the mobility trends of Italians to regions other than those of their own residence when hospitalization is required.

4.1 Results

We set K from 1 to 5 (where K is the number of depth medoids) and the number of repetitions equal to 100. The procedure was applied with respect to both the regions of residence and the region of hospitalization. In both cases the optimal value of K is 4. To determine the mobility trends, one can look at the heatmap. The adopted color scale ranges from red to green, where red represents a large distance between the clusters, yellow a medium proximity and green a high proximity. Below is the heatmap depicted in Figure 1, which was obtained by placing the clusters of the regions of hospitalization on the rows and the clusters of the regions of residence on

the columns.

One can see that the third cluster of the hospitalization regions is the one with the greatest attractivity, while the fourth cluster shows a lower attractivity. We can also note that regions of residence belonging to the third cluster are those characterized by higher level of mobility while and those belonging to the first cluster show a lower mobility degree.

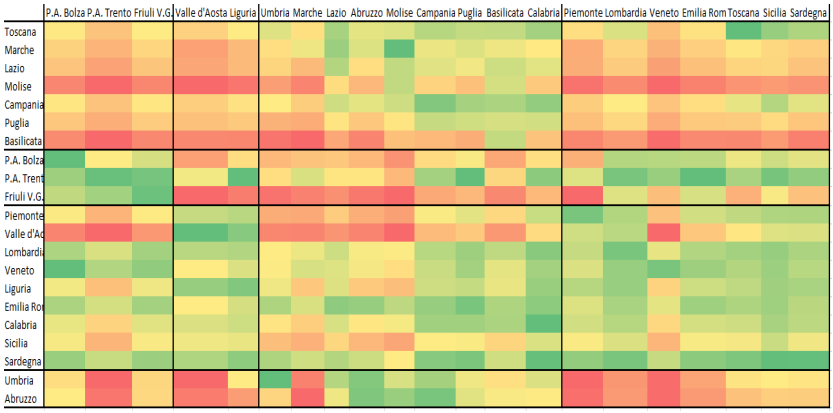


Fig. 1 Heatmap and spherical cluster analysis of Italian interregional healthcare mobility in 2017.

5 Conclusions

We propose an alternative non-parametric clustering procedure which exploits the notion of data depth for directional data for the cluster analysis of inter-regional healthcare mobility in Italy in 2017. This is possible by applying a simple square root transformation to compositional data. The proposed method appears to be able to provide relevant information. It highlights the tendency to mobility expressed by groups and which regions have a greater (or smaller) attractivity for other regions' residents. In addition, it allows to identify those individuals who generally show a greater or smaller tendency to move to other regions.

References

1. Liu, R.Y., Singh, K.: Ordering directional data: concepts of data depth on circles and spheres. The Annals of Statistics **20**, 1468–1484 (1992)

2. Pandolfo, G., Paindaveine, D., Porzio, G. C.: Distance-based depths for directional data. Canadian Journal of Statistics **46**, 593–609 (2018)
3. Pandolfo, G., D'Ambrosio, A.: Depth-based classification of directional data. Expert Systems with Applications **169**, 1–8 (2021)
4. Stephens, M.A.: Use of the von Mises distribution to analyse continuous proportions. Biometrika **69**, 197–203 (1982)

Automatic Fake News Detection to Ensure Quality of News Articles

Rilevamento automatico delle Fake News per garantire la qualit degli articoli di notizie

Bonaventure F.P. Dossou and Adalbert F.X. Wilhelm

Abstract Access to information is an inherent right to every living human being. Defined to be factual information published in newspapers or broadcasted on radio or television, news helps us daily to be aware of events going on in our societies and all over the world. Recent years have seen a fast-growing rate of fake news, conducting to disinformation. In NLP, the battle to use the power of AI to detect, reduce and eradicate the propagation of fake news is an ongoing trend. In this report, using the ISOT fake news dataset, we test the importance of choosing the right embedding model. We integrate the most efficient embedding model, and we implement a reinforcement learning framework to perform binary classification of news articles. In test runs the framework achieves an accuracy of classification of 99.90%.

Abstract *L'accesso all'informazione un diritto inerente ad ogni essere umano vivente. Definite come informazioni fattuali pubblicate sui giornali o trasmesse alla radio o alla televisione, le notizie ci aiutano quotidianamente ad essere consapevoli degli eventi che accadono nelle nostre societ e in tutto il mondo. Gli ultimi anni hanno visto una rapida crescita del tasso di fake news, conducendo alla disinformazione. In NLP, la battaglia per utilizzare la potenza dell'IA per rilevare, ridurre e sradicare la propagazione delle fake news una tendenza in corso. In questa relazione, utilizzando il dataset di fake news ISOT, testiamo l'importanza di scegliere il giusto modello di embedding. Integriamo il modello di incorporazione pi efficiente e implementiamo un framework di apprendimento di rinforzo per eseguire la classificazione binaria degli articoli di notizie. Nei test il framework raggiunge un'accuratezza di classificazione del 99,90%..*

Key words: text classification, reinforcement learning, word embedding

Bonaventure F.P. Dossou
Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, e-mail: b.dossou@jacobs-university.de

Adalbert F.X. Wilhelm
Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, e-mail: a.wilhelm@jacobs-university.de

1 Introduction

The first two decades in the 21st century saw a rapid emergence of the internet and social networks which became an integral part of our daily lives sharing our days, conversing with our friends and families, and sharing our work, to name a few. This implies the generation of an enormous amount of data daily. Unfortunately, the improper handling of these data has given rise to very deplorable situations, facilitating the spread of fake news. The automatic detection of fake news, their reduction and eradication is therefore a great challenge, which remains relevant today. Quite some research has been done in this regard, using many learning techniques such as supervised learning, unsupervised learning, reinforcement learning, or the arising self-supervised learning. Despite these efforts, there is still room for exploration, to leverage the efficiency of reinforcement learning in the fake news detection open challenge.

Reinforcement learning (RL) is one of the learning concepts of Machine Learning (ML). The main goal is to teach an intelligent agent, in a given environment, how to take actions in order to maximize the reward function. RL has many successful stories including its successful application in robot control, telecommunications, checkers, and Go (AlphaGo). Our main contribution in this paper is to demonstrate the ability of an RL agent to successfully classify the fakes news from the real news.

Section 2 will cover a brief overview on initiatives for fake news detections and the use of reinforcement learning with similar tasks. In section 3 we will describe the RL framework and in section 4 we present the ISOT dataset and give the results of our RL implementation for it.

2 Related Works

Algorithmic development for fake news detection has been a growing research field in the past few years, particularly fuelled by the American political debate. [Wang, 2017] introduced a benchmark dataset called "Liar", which has been made of more than tenthousand carefully and manually labeled short sentences over numerous contexts taken from https://www.politifact.com/. On this dataset, the authors built a ConvNet + biLSTM model to integrate meta-data with text. The results showed an improvement in a text-only deep learning model with an accuracy of 27%. [Yang et al., 2018] introduced a similar CNN-based model for fake news detection, which showed effectiveness in the news detection.

[Monti et al., 2019] introduced a new model for automatic fake news detection using geometric deep learning. The proposed model is a mere generalization of Convolutional Neural Networks (CNNs) to graphs notions. The model was trained and tested on Twitter news stories and achieved a score of 92.7% ROC AUC.

On the RL approach point of view, [Zhang et al., 2018] showed how to extract a better and structured representation for text classification. Authors proposed a RL method to learn sentence representation by discovering optimized structures

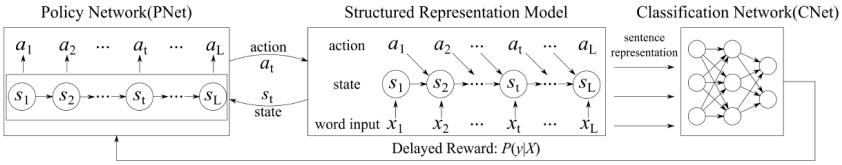


Fig. 1 Overall framework structure, taken from [Zhang et al., 2018]

automatically. The paper introduces two models: Information Distilled LSTM (ID-LSTM) and Hierarchically Structured LSTM (HS-LSTM) which demonstrated better performances on benchmark datasets, compared to previous state-of-the-art models introduced on the same datasets.

Similarly to [Zhang et al., 2018], [Wang et al., 2019] proposed a reinforced weakly-supervised framework that leverages users' reports to enlarge the amount of training data for the detection of fake news. Called WeFEND, the framework consists of three parts: an annotator, a reinforced selector, and a fake news detector. Authors claimed that the annotator can automatically assign weak labels for unlabeled news based on users' reports. As far as the reinforced selector is concerned, using reinforcement learning techniques it chooses high-quality samples from the weakly labeled data and filters out the low-quality samples that could reduce the detector's prediction performance.

3 Reinforcement Learning Framework

Our RL implementation is inspired from the framework proposed by [Zhang et al., 2018], as we converted the detection task to a classification task. As shown in Figure 1, the framework is made of three main parts: the policy network (PNet), the structured representation model (SRM), and the classfication network (CNet).

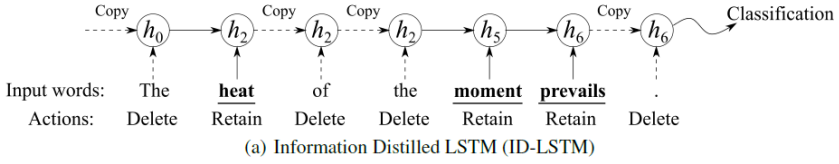
PNet (hereafter called Π) is based on a stochastic policy

$$\pi(a_t|s_t,\theta) = \sigma(W * s_t + b)$$

where $\sigma = sigmoid(x)$, $\theta = (W,b)$ with W and b being respectively weights and biases matrices.

During inference, the action with the maximal probability $a_t^* = argmax(\pi(a_t|s_t,\theta))$ is chosen in order to obtain better prediction. π uses a delayed reward to guide the policy learning. The states encode the current input and previous contexts. At each state s_t, π samples an action a_t with its probability, until the end of a sentence, producing a sequence of action for the sentence s.

After the sampling of all actions by Π, the structured representation of a sentence is determined the SRM, and passed to CNet to obtain $P(y|x)$ where y is the label. The predicted distribution $P(y|x)$, will be used to calculate the reward r_s



(a) Information Distilled LSTM (ID-LSTM)

Fig. 2 ID-LSTM Structure, taken from [Zhang et al., 2018]

The parameters θ of Π are optimized using the reinforce algorithm proposed by [Williams, 1992], and the policy gradients [Sutton et al., 1999]. The objective function is defined as follow:

$$J(\theta) = E(s_t, a_t)\ P_{\theta(s_t,a_t)} r(s_1 a_1 ... s_m a_m)$$

$$= \sum_{s_1 a_1 ... s_m a_m} \prod_t \pi_\theta(a_t|s_t) r_m$$

The reward described above is calculated for a unique sample $X = x_1 x_2 ... x_m$. However, since the state at step $t+1$ is fully determined by the state and action at step t, the probabilities $p(s_1)$ and $p(s_{t+1}|s_t, a_t)$ are equal to 1. The parameters of the network and the Π are updated with the following gradient:

$$\nabla_\theta J(\theta) = \sum_{t=1}^{m} r_m \nabla_\theta log \pi_\theta(a_t|s_t)$$

In our report we used the Information Distilled LSTM (ID-LSTM). The main idea is to build sentences representations by distilling most important words and removing irrelevant words in a sentence. The authors argued that this will allow to learn more efficiently task-relevant representations for classification. For example, words like 'to', 'a', 'the' and other articles are less likely to be relevant and important to the sentence meaning. Hence, using ID-LSTM (Figure 2), the final representation can be purified and condensed for a better and suitable classification.

ID-LSTM translates the actions obtained from Π to a structured representation of a sentence. Formally, given a sentence $X = x_1 x_2 ... x_m$, there is a corresponding action sequence $A = a_1 a_2 ... a_m$ obtained from Π. Each action a_i of the word x_i is chosen from $\{retain, delete\}$, where *retain* indicates that the word is retained in a sentence, and *delete* means that the word is deleted and it has no contribution to the final sentence representation. To make classification, the last hidden state of ID-LSTM is taken as input to the classification network (CNet) $p(y|x) = softmax(W_s h_m + b_s)$. The reward r_m is computed, taking the logarithm of the predicted probability, and adding the ratio of deleted words (m') in the final sentence representation over the total initial number of words of the initial sentence (m):

$$r_m = log(p(y|x)) + \gamma \frac{m'}{m}$$

$\gamma \in [0,1]$ is a trade-off hyper-parameter.

The classification network outputs the probability distribution over class labels (real, fake) based on the structured representation obtained from ID-LSTM. The CNet is trained using the categorical cross entropy loss, where K is the number of categories. In our case $K = 2$, which gives the following adapted loss function

$$\mathscr{L} = \sum_{x \in \mathscr{D}} - \sum_{y=1}^{2} \hat{p}(y,x) log(p(y|x))$$

where $\hat{p}$ is the expected distribution, and p the predicted one.

4 Data, Experiments and Results

The ISOT fake news dataset introduced in [Ahmed et al., 2017], by the University of Victoria, contains two types of articles: fake and real news. The real news of the dataset were obtained by crawling articles from https://www.reuters.com/. The fake news were collected from different sources including by https://www.politifact.com/ and Wikipedia. The dataset is made of 21417 real news articles, and 23481 fake news articles, cleaned and labeled by experts. The raw vocabulary size of our dataset is 104k, with the largest sentence length being 4953. This yields in a vector space of size $(104k, 4953)$: this computationally expensive. To reduce the the complexity of the space, we filtered articles by length with a reasonable threshold of 150 (i.e we selected only articles whose length ≤ 150). Alternatively, we set the embedding vector space length to 300. We then have a vector space of size $(35984, 300)$: this is still computationally expensive, but already better than the initial and raw size. We split the dataset into three parts: training 80%, validation 10%, and testing 10%. We ran the experiments on 30 epochs. We also implemented the *EarlyStop* method, with a *patience* of 10 to control the under/over fitting of the framework (of the classifier to be more precise). Every component of the framework, has been trained with a learning rate of 0.003, and a *momentum* of 0.9, as we used Stochastic Gradient Descent (SGD) [Robbins, 2007] as optimizer, with Dropout [Srivastava et al., 2014] as regularization. The full training was done in two steps:

1. we trained firstly the classifier, which is a LSTM model. The classifier after 30 epochs achieved an accuracy of 99.69%, as shown in the classification report of the classifier
2. secondly, the pre-trained classifier model is used as RL agent, and trained in the defined environment, with the policy network Π. The RL agent achieved an accuracy score of 99.90%, an almost-perfect score.

After the full training, we tried to leverage the RL agent performance without the pre-trained classifier. We trained the RL agent directly in its environment, and the

accuracy of prediction dropped to 75%, with **no correct** prediction on the fake news articles. This means that pre-training a-priori the classifier helped the RL agent to produce significant rewards toward a successful classification.

5 Conclusion

Throughout this report, we demonstrated the abilities of different embedding models to effectively represent words, sentences and their contexts. We showed how their representations could improve the system performance on the task at hand. Adapting the implementation of [Zhang et al., 2018] we built a reinforcement learning framework on a subset of the initial dataset (described in section **??**), to perform the fake news detection task that we conveyed to a binary classification task. The reinforcement learning agent achieved an accuracy of classification of 99.90%. We corroborate that this result, albeit on a small dataset, shows the efficiency of RL in the fake news detection task.

References

[Ahmed et al., 2017] Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. pages 127–138.

[Monti et al., 2019] Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning.

[Robbins, 2007] Robbins, H. (2007). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):19291958.

[Sutton et al., 1999] Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 10571063, Cambridge, MA, USA. MIT Press.

[Wang, 2017] Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

[Wang et al., 2019] Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., and Gao, J. (2019). Weak supervision for fake news detection via reinforcement learning.

[Williams, 1992] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(34):229256.

[Yang et al., 2018] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., and Yu, P. S. (2018). Ti-cnn: Convolutional neural networks for fake news detection.

[Zhang et al., 2018] Zhang, T., Huang, M., and Zhao, L. (2018). Learning structured representation for text classification via reinforcement learning.

Session of solicited contributes SS11 – *Multi-way Methods for Evaluation Service*
Organizer and Chair: Michele Gallo

Multiple factor analysis with external information on PISA survey data

Analisi Fattoriale Multipla con Informazioni Esterne sui dati dell'indagine PISA

Violetta Simonacci, Marina Marino, Maria Gabriella Grassia and Michele Gallo

Abstract OECD-PISA survey data include performance measurements, expressed as tables of plausible values, and a variety of socio-biographical information. Such data, if properly modeled, can provide useful insights on the causes of low performing students. A methodology which deals with multivariate sets of plausible values and investigates the effects of context variables without assumptions is required. Here Multiple Factor Analysis with External Information is proposed. Specifically, after defining context variable groupings, a partitioning of the variability structure of the data tables is carried out using projection operators than a simplified Multiple Factor Analysis with bootstrap is performed.

Abstract *I dati delle indagini OCSE-PISA includono misurazioni della performance, espresse come insiemi di valori plausibili, e una varietà di informazioni socio-biografiche. Tali dati, se opportunamente modellati, possono fornire utili spunti sulle cause dello scarso rendimento degli studenti. Si rende necessaria dunque una metodologia che si occupi di insiemi multivariati di valori plausibili e indaghi gli effetti delle variabili di contesto. Qui è proposta l'Analisi Fattoriale Multipla con Informazioni Esterne. In particolare, dopo aver definito i raggruppamenti delle variabili di contesto, è eseguito un partizionamento della struttura di variabilità delle tabelle mediante operatori di proiezione, poi è eseguita un'Analisi Fattoriale Multipla semplificata con bootstrap.*

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

Key words: bootstrap replicates, context variables, MFA, multiset data, plausible values

1 Introduction

The periodic assessment of 15 year old students carried out with the OECD's Programme for International Student Assessment (PISA) reveals a considerable performance disparity among Italian regions. The Campania Region Administration has been actively promoting research programs to investigate the causes behind its low performance.With this purpose, their database of PISA results was made available for exploratory analysis.

The data display a complex structure which includes performance levels in different domains expressed as sets of plausible values and socio-economic, attitude and cultural measures. To provide a useful insight into students performance, a fitting methodology is outlined. Modeling tools were chosen to address two specific research goals: i) studying the multivariate structure of performance by correctly dealing with tables of plausible values; and ii) assessing group-level differences on the basis of context variables, without major assumptions.

To avoid bias estimation of group-level differences PISA performance measurements are not provided as single point estimates for individuals, but rather as ten plausible values randomly drawn for the posterior distributions. Randomly extracted instances are thus provided for all individuals in each domain, yielding ten fully crossed "*student by domain*" tables. Modeling single or averaged plausible values does not take into account uncertainty of sampling or testing unreliability [6]. Any analysis on such data should be carried out on each table separately and, only afterwards, results can be aggregated.

Given these considerations, a bilinear exploratory method which decomposes each table separately and then provides a compromise solution would be a good choice. In this perspective an adaptation of Multiple Factor Analysis (MFA) is proposed. Bootstrap replicates are included in the model to deal with sampling variance.

After choosing an appropriate exploratory model, the second research goal can be addressed. To understand the causes of the performance gap on the basis of the background variables, group differences should be evaluated within the model. This is achieved by extending the methodology of Principal Component Analysis (PCA) with External Information [5] to MFA. This method consists in segmenting the total information of the data into two structures of variability, one explained by the selected background variable(s) (external analysis) and a residual one (internal analysis). Separate PCA are carried out on each variability structure of interest. The adaptation of this method to MFA is straight forward.

To sum up, the aim of this work is to assess the effects of background variables on the multivariate structure of Campania Region PISA results. To do so, an MFA with bootstrap resampling and External Information is carried out. In Section 2 the

database in described in detail; in Section 3 the methodology is outlined and in Section 4 some conclusive remarks and preliminary results are presented.

2 Performance Data: PISA 2018 in Campania

The Campania Region Administration provided a large dataset of PISA responses and measures referring to the 2018 survey. The evaluation was carried out on a sample of 1670 individuals. After excluding students with multiple missing entries, the sample is reduced to 1548 units. The data can be subdivided into performance variables and background information.

Let us focus on performance measures first. In 2018, five domains were assessed: "Problem solving" (Pr _ Sol), "Financial Literacy" (Fin _ L) and the three core domains, "Reading" (*RDN*), "Mathematics" (*MAT*), "Science" (*SCI*). Ten plausible values were imputed for each domain and included in the dataset. In 2018 all domains were scaled in the same way and each set of plausible values was drawn at the same time.

Such data can be arranged into ten tables each of size (1548×5), yielding a two-way matrix for each plausible value imputation. An output similar to a repeated measures design is generated. It is clear that a suitable analysis of such data requires a tool which models performance while taking into account sampling uncertainty by considering all sets of plausible values in the solution. A methodology based on MFA is proposed here and briefly introduced in the following section.

In addition to performance measurements, numerous context variables are also provided for each student. The first step in selecting background variables of interest is to eliminate all the measures with a large amount of missing values and redundant information (items already included in other estimates). A total of 30 variables is selected. Given the large amount of information, after a quick first assessment of significance, only the most relevant results will be presented.

Most of these variables were built as latent constructs based on the responses to multiple items and expressed as Weighted Likelihood Estimates (WLE) on an interval-scale. In the External Information analysis, these quantitative measurements are transformed into dummies to build homogeneous student groupings. Seven groups are identified for each WLE variable based on cut-off values of the index score. The following groups are constructed: "very low score", "low score", "score below the average", "average", "score above the average", "high score", "very high score". Such classification is possible because the variables are standardized with respect to the distribution of OECD countries to have a mean of 0 and a standard deviation of 1. Consequently, a negative score does not mean that a student has answered negatively to a question (or set of questions) but simply that they answered less positively than the average student in OECD countries.

3 Methodology

3.1 MFA for plausible values tables

Multiple Factor Analysis is a technique based on singular value decomposition (SVD) designed for the analysis of $[1,\dots,k,\dots,K]$ tables $\mathbf{X}_k$ which collect sets of variables on the same $[1,\dots,i,\dots,I]$ observations [2, 3]. Generally the K groups of variables differ from one another and each $\mathbf{X}_k$ stores its own $[1,\dots,j_k,\dots,J_K]$ variables. In some instances, however, all $\mathbf{X}_k$ may refer to the same variables $[1,\dots,j,\dots,J]$ measured under different conditions (repeated measures). In the PISA dataset this latter simplified version of MFA is considered but, instead of repeated measures, the tables contain different plausible values imputations.

The procedure is outlined as follows. First, each table is decomposed by (truncated) SVD and scaled by dividing its elements by the first singular value. These scaled tables are then joined in a single wide matrix which is also subsequently analyzed by SVD. The results include common scores and loadings, generally known as compromise or consensus, and partial factor scores for each of the K tables. Formally, these subsequent steps are executed.

1. Truncated SVD is performed on each $\mathbf{X}_k$ to retrieve the first singular value

$$\text{SVD}(\mathbf{X}_k) = \mathbf{u}_k \sigma_k \mathbf{v}_k' \qquad with \quad 1,\dots,k,\dots,K \tag{1}$$

   σ_k is the first singular value of $\mathbf{X}_k$; $\mathbf{u}_k(I\times 1)$ and $\mathbf{v}_k(J\times 1)$ are the first left and right singular vectors, respectively.

2. The singular values $\sigma_1,\dots,\sigma_k,\dots,\sigma_K$ are used to build the matrix of weights $\mathbf{A}(K\times K)$:

$$\mathbf{A} = diag(\alpha_1,\dots,\alpha_k,\dots,\alpha_K) \qquad with \quad \alpha_k = \frac{1}{\sigma_k^2} = \sigma_k^{-2} \tag{2}$$

3. Given $\mathbf{X} = [\mathbf{X}_1 \mid \dots \mid \mathbf{X}_k \mid \dots \mid \mathbf{X}_K]$, a weighted wide matrix $\tilde{\mathbf{X}}$ of juxtaposed K tables can be formulated:

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{A} = [\alpha_1\mathbf{X}_1 \mid \dots \mid \alpha_k\mathbf{X}_k \mid \dots \mid \alpha_K\mathbf{X}_K] \tag{3}$$

4. The wide matrix $\tilde{\mathbf{X}}$ is decomposed in R factors by SVD:

$$\text{SVD}(\tilde{\mathbf{X}}) = \tilde{\mathbf{U}}\tilde{\Delta}\tilde{\mathbf{V}}' \tag{4}$$

   where $\tilde{\Delta}(R\times R)$ is the diagonal matrix of singular values while $\tilde{\mathbf{U}}(I\times R)$ and $\tilde{\mathbf{V}}(J\cdot K\times R)$ are the left and right singular vector matrices respectively. The matrix $\tilde{\mathbf{V}}$ can be expressed as:

$$\tilde{\mathbf{V}} = \left[\tilde{\mathbf{V}}_1' | \ldots | \tilde{\mathbf{V}}_k' | \ldots | \tilde{\mathbf{V}}_K'\right] \tag{5}$$

where the generic matrix $\tilde{\mathbf{V}}_k(J \times r)$ stores the right singular vectors of the corresponding matrix $\tilde{\mathbf{X}}_k$.

At this point the consensus solution can be explored. Compromise scores for individuals are easily found by $\mathbf{F} = \tilde{\mathbf{U}}\tilde{\Delta}(I \times R)$. Compromise loadings can be computed only in the special case of repeated measures. Each right singular vector matrix $\tilde{\mathbf{V}}_k$ in eq.5 is scaled back to its original variability structure so that K rescaled $\mathbf{Q}_k$ matrices are yielded with $\mathbf{Q}_k = \frac{1}{\alpha_k}\tilde{\mathbf{V}}_k$.

Now the compromise factor loading matrix $\bar{\mathbf{Q}}(J \times R)$ with element $\bar{q}_{jr} = \frac{1}{K}\sum_{k=1}^{K} q_{jkr}$ can be defined as the barycenter of the partial factor loadings.

Bootstrap resampling is incorporated into the procedure to provide information on sample variability, following PISA technical repots indications [1, 4].

3.2 Adding external information

To study the impact of various context variables on the structure yielded by MFA, the External Information methodology can be added to the model in the following manner. First, for each external variable a generic matrix $\mathbf{G}$ of dimension $(I \times m)$ with $(m < I)$ can be built, where m is the number of groups defined within the variable. The columns of the matrix are dummies which indicate if a subject belongs to a certain group or not.

Successively the variability structure within each generic k-th table can be decomposed as follows in order to study the effect of $\mathbf{G}$:

$$\mathbf{X}_k = \mathbf{G}\mathbf{B}_k + \mathbf{E}_k \tag{6}$$

where $\mathbf{E}_k$ is the residual matrix of dimension $(I \times J_k)$ which includes internal variability (not explained by $\mathbf{G}$) and $\mathbf{B}_k$ contains the coefficient to estimate by minimizing the sum of squares of residuals $SS(\mathbf{E}_k) = tr(\mathbf{E}_k'\mathbf{E}_k)$. Thus, we have $\hat{\mathbf{B}}_k = \mathbf{P}_G + \mathbf{X}_k$ were $\mathbf{P}_G = \mathbf{G}(\mathbf{G}^t\mathbf{G})^{-}\mathbf{G}^t$ is the orthogonal projection operator generated by $\mathbf{G}$. The decomposition can then be rewritten as:

$$\mathbf{X}_k = (\mathbf{P}_G + \mathbf{P}_G^{\perp})\mathbf{X}_k = \mathbf{P}_G\mathbf{X}_k + \mathbf{P}_G^{\perp}\mathbf{X}_k \tag{7}$$

where $\mathbf{P}_G^{\perp}$ is the orthogonal complement of $\mathbf{P}_G$. It is clear that the first term of the equation contains the external information explained by $\mathbf{G}$ and the second term only the residual one.

In light of these steps, eq. 1 and eq. 3 of the MFA procedure can be modified to focus on external information. Specifically all $\mathbf{X}_k$ tables in the formula must be pre-multiplied by $\mathbf{P}_G$. MFA steps can then be carried out as specified in the previous section.

4 Preliminary considerations

PISA surveys data represent an enormous source of information not only on students' performance, but also on their background with respect to school, family and society. Identifying modeling tools suitable for the complexity of their structure is an ongoing challenge.

The proposed methodology, based on MFA with external information, has been chosen to adequately deal with plausible values tables and to evaluate students' results in relation to their context.

Preliminary results have highlighted interesting effects of the selected external variables on the performance of students. Some of the most interesting context variables are "Sense of anxiety", "Sense of belonging", "Emotional support of parents" and "Family wealth". In reference to the latter measure, for example, it was observed how the first axis explains most of the external variability (more than 80%). Groups of students with low scores record lower levels of performance in all domains. As the "Family wealth" score increases the level of performance also increases. However, this linear trend disappears at the highest level.

Results will be displayed with the support of the symmetrical MFA biplot representation with confidence interval ellipses. Some methodological remarks will also be discussed on the use of projection operators in MFA, specifically on how the choice of the procedure step in which the partition of variability is introduced impacts results.

References

1. Babamoradi, H., van den Berg, F., Rinnan, Å.: Bootstrap based confidence limits in principal component analysis—a case study. Chemometrics and Intelligent Laboratory Systems **120**, 97–105 (2013)
2. Escofier, B., Pages, J.: Multiple factor analysis (afmult package). Computational statistics & data analysis **18**(1), 121–140 (1994)
3. Pagès, J.: Multiple factor analysis by example using R. CRC Press (2014)
4. Pagés, J., Husson, F.: Multiple factor analysis with confidence ellipses: a methodology to study the relationships between sensory and instrumental data. Journal of Chemometrics: A Journal of the Chemometrics Society **19**(3), 138–144 (2005)
5. Takane, Y., Shibayama, T.: Principal component analysis with external information on both subjects and variables. Psychometrika **56**(1), 97–120 (1991)
6. Von Davier, M., Gonzalez, E., Mislevy, R.: What are plausible values and why are they useful. IERI monograph series **2**, 9–36 (2009)

A three-way analysis of well-being in Italy over time

Un'analisi a tre-vie del benessere in Italia nel tempo

Laura Bocci and Donatella Vicari

Abstract In this paper, we provide an analysis of the Italian regional well-being according to the equitable and sustainable well-being (BES) indicators pertaining to work and social domains in the years 2010-2020, in order to identify differences and similarities across years. A three-way approach is adopted by using the Extended STATIS method, an extension of the standard STATIS method, in order to capture the different role of the well-being indicators over time.

Abstract *In questo lavoro viene analizzato il benessere regionale italiano secondo gli indicatori di benessere equo e sostenibile (BES) relativi agli ambiti del lavoro e sociale negli anni 2010-2020 al fine di identificare differenze e somiglianze nel tempo. Si segue un approccio a tre-vie attraverso il metodo STATIS Esteso che rappresenta un'estensione del metodo STATIS, allo scopo di catturare il differente ruolo degli indicatori del benessere nel tempo.*

Key words: Well-being indicators, Italian regions, Extended STATIS

1 Introduction

For many decades, the Gross Domestic Product (GDP) has been considered the most relevant indicator of a country's progress, but experience has shown that it is lacking in accurately reflecting the progress of society towards broad-based prosperity. The final report of the European Commission for Measuring Economic Performance and Social Progress, known as "Stiglitz report" (Stiglitz et al., 2009), defines the guidelines for measuring well-being by adopting a multidimensional approach that goes "Beyond GDP".

[1] Laura Bocci, Department of Economic and Social Sciences, Sapienza University of Rome, Rome, Italy; email: laura.bocci@uniroma1.it

Donatella Vicari, Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy; email: donatella.vicari@uniroma1.it

In 2010, Istat (Italian Institute of Statistics) and Cnel (Italian Council for Economics and Labour) launched a project to measure equitable and sustainable well-being (Benessere Equo-Sostenibile - BES) that aims at evaluating the progress of society not only from an economic, but also from a social and environmental point of view. Hence, BES follows a multidimensional approach to measure well-being, in order to complement the indicators related to production and economic activity with measures of the key dimensions of well-being, together with measures of inequality and sustainability. Grounded on these considerations, since 2013, the various facets of well-being declined in 12 domains have been measured by using several indicators selected to take into account the two basic concepts of equity and sustainability and specifically designed to capture economic, social and environmental sustainability (Burchi and Gnesi, 2016). In order to facilitate the understanding of the evolution of the different dimensions of well-being, a composite index is computed for each domain by only few indicators from the whole set which are then aggregated on a yearly basis by means of an aggregative-compensative method (Mazziotta and Pareto, 2016). Since 2013, the indicators together with composite indices are published annually in the Bes Report to raise awareness of the strengths and difficulties of the Italian regions.

Differences and similarities in the performance of the Italian regions in each domain of well-being over time can be investigated. In this framework, this study focuses on two out of 12 domains: Work and life balance and Social relationships. A decent job adequately paid is a key element contributing to people's well-being and lack of quality employment has a negative impact on the level of well-being. Moreover, a fundamental part of people's social capital affecting individual well-being is represented by relational networks, especially family and friendships.

Therefore, the objective of this study is a) to analyse the relationships between the indicators by explicitly taking into account their different role over time and b) to obtain a ranking of the Italian regions over time. In order to achieve these goals, the two sets of indicators used for computing the composite indices of Work and life balance domain and Social relationships domain are considered. All the indicators are measured on the 20 Italian regions from 2010 to 2020. Data are analysed by adopting a three-way approach through the Extended STATIS method (Bocci and Vicari, 2021). The Extended STATIS, which is an extension of the standard STATIS (acronym for the French expression "Structuration des Tableaux à Trois Indices de la Statistique" - Escoufier, 1980), is assumed as a generalization of the Principal Component Analysis (PCA) for three-way data.

The paper is organized as follows. Section 2 contains the data description and the methodology used in the analysis, while the main results are shown in Section 3.

2 Data and methods

Two of the 12 domains relevant to the Italian regional well-being are taken into consideration: Work and life balance (WRB) and Social relationships (SR). Each of the two data sets is a three-way three-mode data set pertaining to three different sets

of entities (i.e., units, variables and occasions). Specifically, the units correspond to the $I = 20$ Italian regions, a selection of the BES indicators for each domain represents the J variables (either J=6 for Work and life balance domain or J=8 for Social relationships domain) and the occasions are $K = 11$ years from 2010 to 2020.

Therefore, for each of the two domains under study, the available information consists of K=11 data matrices $\mathbf{X}_k$ of size ($I \times J$) containing the values of J (either J=6 or J=8) indicators collected on the same I=20 Italian regions in year k (k = 1,...,11).

The indicators considered represent a selection of those analysed within the BES 2020 Report (Istat, 2021) and used by Istat in the construction of the composite index for the domain themselves.

The indicators from the Work and life balance domain are: Employment rate (WLB1), Share of employed persons with temporary jobs for at least 5 years (WLB2), Share of employees with below 2/3 of median hourly earnings (WLB3), Share of employed persons not in regular occupation (WLB4), Share of employed persons who feel satisfied with their work (WLB5), Involuntary part time (WLB6).

The indicators from the Social relationships domain are: Satisfaction with family relations (SR1), Satisfaction with friends relations (SR2), People to rely on (SR3), Social participation (SR4), Civic and political participation (SR5), Voluntary activity (SR6), Association funding (SR7), Generalized trust (SR8).

The choice of these indicators is based on a twofold criterion: on the one hand, the ability to represent the different characteristics of each domain and, on the other hand, the availability of data at the regional level over time.

The indicators have been adjusted according to their polarity, i.e. the sign of the relationship between each indicator and the phenomenon to be measured, and also standardized to zero mean and standard deviation equal to 1. Due to the polarity adjustment, a positive value of each indicator means a positive condition: for example, a positive value of "Share of employed persons not in regular occupation" indicator means that the share of people employed in non-regular occupation is below average.

Let $\underline{\mathbf{X}} = [\mathbf{X}_1, \dots, \mathbf{X}_k, \dots, \mathbf{X}_K]$ be the $(20 \times JK)$ matrix formed by collecting the K=11 matrices $\mathbf{X}_k$ next to each other and $\mathbf{Z}$ be the $(400 \times JK)$ matrix formed by the (column-wise) Khatri–Rao product of $\underline{\mathbf{X}}$ with itself, i.e., $\mathbf{Z} = \underline{\mathbf{X}} | \otimes | \underline{\mathbf{X}} = (\mathbf{x}_{1k} \otimes \mathbf{x}_{1k}, \dots, \mathbf{x}_{jk} \otimes \mathbf{x}_{jk}, \dots, \mathbf{x}_{JK} \otimes \mathbf{x}_{JK})$ where $\mathbf{x}_{jk}$ is the j-th column of $\mathbf{X}_k$ ($k = 1, \dots, 11$) and $\otimes$ denotes the Kronecker product.

Given the domain, the Extended STATIS aims to analyse the structure of the data array performing two main steps:

1) the *inter-structure* analysis, which consists in deriving two optimal sets of weights, for the years and for the indicators, from the analysis of the similarities between years; the two sets of weights are then used to get an optimal consensus (the so-called *compromise*) as representative as possible of the 11 data matrices (years);
2) the *intra-structure* analysis, where a generalized Principal Component Analysis (PCA) of the compromise is performed to obtain the best representation of the 20 regions in a space common to all years.

The goal of the first step of the Extended STATIS method is to estimate the two sets of weights $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k, \dots, \alpha_{11})'$ for the K=11 years and $\mathbf{w} = (w_1, \dots, w_j, \dots, w_J)'$ for the J (either J=6 or J=8) indicators in order to optimize the similarities between the K=11 matrices $\mathbf{X}_k$ and the *compromise* $\mathbf{S}_+$ defined as

$$\mathbf{S}_+ = \underline{\mathbf{X}}(\mathbf{A} \otimes \mathbf{W})\underline{\mathbf{X}}' \quad (1)$$

where **A** and **W** are two diagonal matrices whose main diagonal elements are $\boldsymbol{\alpha}$ and **w**, respectively.

The two sets of weights $\boldsymbol{\alpha}$ and **w** can be estimated by solving the following least-squares problem

$$\max_{\boldsymbol{\alpha},\mathbf{w}} g(\boldsymbol{\alpha},\mathbf{w}) = \|\mathbf{S}_+\|^2 \quad \text{subject to} \quad \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1 \quad \text{and} \quad \mathbf{w}'\mathbf{w} = 1 \quad (2)$$

which consists in the maximization of the variance of the compromise $\mathbf{S}_+$. Actually, problem (2) can be reformulated as a constrained regression problem as follows

$$\max_{\boldsymbol{\alpha},\mathbf{w}} g(\boldsymbol{\alpha},\mathbf{w}) = \|\mathbf{Z}(\boldsymbol{\alpha} \otimes \mathbf{w})\|^2 \quad \text{subject to} \quad \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1 \quad \text{and} \quad \mathbf{w}'\mathbf{w} = 1. \quad (3)$$

An Alternating Least-Squares (ALS) algorithm is proposed to solve problem (2)-(3) where two steps are alternated and iterated until convergence. Specifically, each step estimates in turn:

a) the vector of the year weights $\boldsymbol{\alpha}$, given **w**, by maximizing $\boldsymbol{\alpha}'\mathbf{M}\boldsymbol{\alpha}$ subject to $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, where $\mathbf{M} = (\mathbf{I}_{11} \otimes \mathbf{w})'\mathbf{Z}'\mathbf{Z}(\mathbf{I}_{11} \otimes \mathbf{w})$ and $\mathbf{I}_{11}$ is an identity matrix of order 11;
b) the vector of the indicator weights **w**, given $\boldsymbol{\alpha}$, by maximizing $\mathbf{w}'\mathbf{N}\mathbf{w}$ subject to $\mathbf{w}'\mathbf{w} = 1$, where $\mathbf{N} = (\boldsymbol{\alpha} \otimes \mathbf{I}_J)'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\alpha} \otimes \mathbf{I}_J)$ and $\mathbf{I}_J$ is an identity matrix of order *J*.

The solutions of the two steps a) and b) of the algorithm are achieved by taking the first eigenvector of **M** and **N**, respectively.

In the second step (the intra-structure analysis) of the method, the structure of the relationships between the Italian regions is investigated by performing a PCA of the compromise $\mathbf{S}_+$ in (1) to analyse the space spanned by the first principal components which will be referred to as compromise space.

3 Results

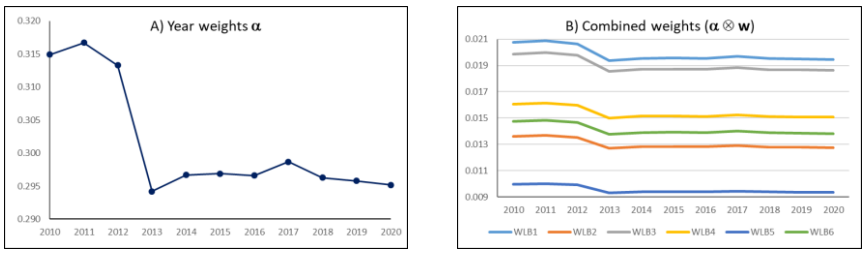
The Extended STATIS method was applied separately to two domains: the (20×6×11) matrix of the Work and life balance domain and the matrix of order (20×8×11) of the Social relationships domain. The objective was to explore the differences and similarities of the Italian regions with respect to the two well-being domains, taking into account the different role of the indicators over time and then to define the position of the regions across years.

At first, the inter-structure step is performed by deriving the optimal weights $\boldsymbol{\alpha}$ and **w** for each domain. The compromise $\mathbf{S}_+$ accounts for about the 61% of the total variability of the indicators across years for both domains.

Figure 1 shows the optimal weights of the years and the combined weights $(\boldsymbol{\alpha} \otimes \mathbf{w})$ for the Work and life balance domain. The pattern of $\boldsymbol{\alpha}$ weights highlights the similarity structure across years (Figure 1A). Only the three most similar years from 2010 to 2012 have high weights, while lower weights are assigned to the remaining years. Moreover, the two least similar years 2013 and 2020 have the lowest

weight likely due to the consequences of the 2011 crisis and the COVID-19 pandemic, respectively. Actually, the evolution of the the Work and life balance domain followed that of the economic cycle showing a sharp fall in 2013, a gradual recover until 2019 and then suffering the negative effects of the COVID-19 pandemic in 2020.

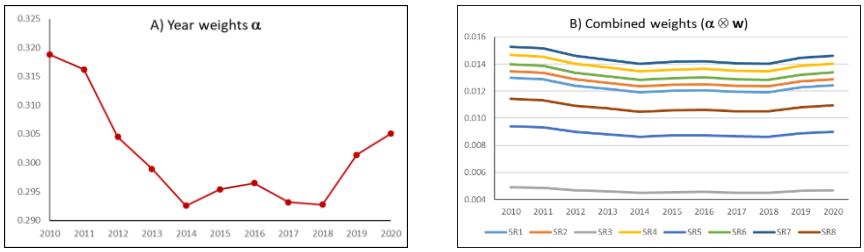
Figure 1: Year weights (A) and combined weights (B) for Work and life balance domain



However, there is no uniformity in the trends of the basic indicators over the period in 2010-2020. In 2014 the "Employment rate" (WLB1) starts to rise again but in 2020 it drops down again due to the COVID-19 pandemic which interrupts the upward trend. On the other hand, the "share of precarious work" (WLB2) and "workers with low pay" (WLB3) improve slightly in 2016 after the downward phase that begins in 2009 and stabilizes in 2018-2019. On the contrary, "irregular work" (WLB4) and the "share of involuntary part-time workers" (WLB6) have continued to worsen since 2010. Looking at the combined weights in Figure 1B, it is possible to evaluate the contribution of the indicators to the similarity structure of the Work and life balance domain in the 11 years. The "Employment rate" (WLB1) and the "Share of employees with below 2/3 of median hourly earnings" (WLB3) have the largest weights followed by the indicators on the stability of work (WLB4, WLB6 and WLB2), while the "satisfaction with the work" (WLB5) contributes to a lesser extent.

Figure 2 shows the optimal year weights **α** and the combined weights (**α ⊗ w)** for the Social relationships domain. Here, years 2010 and 2011 are the most similar with the largest weights, while the years from 2014 to 2018 have lower weights (Figure 2A).

Figure 2: Year weights (A) and combined weights (B) for Social relationships domain



This negative trend has reversed over the last two years (2019-2020). This pattern is due to the fact that "satisfaction with family" (SR1), "friends" (SR2), "relatives and neighbours" (SR3) show a substantial stability in the 2014-2018 period followed by a slight decrease in 2019-2020. "Voluntary activity" (SR6) remains stable over time,

while the "trust in others" (SR8) stabilizes in 2019-2020 after the decrease registered in 2017. On the contrary, in the last two years, "Association funding" (SR7) and "Civic and political participation" (SR5) increase, "Social participation" (SR4) remains stable after a slight increase in 2018. The indicators that most contribute to the similarity structure over time are: "Association funding", "Social participation" and "Voluntary activity" while "People to rely on" contributes least of all (Figure 2B).

The analysis of the compromise $\mathbf{S}_+$ in (1) computed with the optimal weights $\boldsymbol{\alpha}$ and $\mathbf{w}$ for each domain has been carried out in the following intra-structure step of the Extended STATIS method through a PCA. For both domains, the first component of the compromise accounts for a high percentage of variance (81.6% for Work and life balance and 80.6% for Social relationships) and it is highly correlated with all the indicators in each domain defining a well-being domain index over time.

The ranking of the Italian regions in each domain is reported in Table 1. The geography of well-being traced by the two rankings reflects the Italian structural territorial gap, with the regions of northern Italy showing higher levels of well-being than the central and southern regions. In the South, all forms of relational networks seem weaker than in the rest of Italy. Therefore, the greater difficulties in the field of work experienced by the population of the South compared to that of the North are not partially compensated by family and friendships.

Table 1: Ranking of the Italian regions from the first component of the compromise for Work and life balance (WLB) domain and Social relationships (SR) domain

Region	***WLB Rank***	***SR Rank***	***Region***	***WLB Rank***	***SR Rank***
Trentino-Alto Adige	1	1	Umbria	11	10
Valle d'Aosta	2	4	Abruzzo	12	14
Veneto	3	3	Lazio	13	13
Lombardia	4	6	Molise	14	16
Emilia-Romagna	5	5	Sardegna	15	11
Friuli-Venezia Giulia	6	2	Basilicata	16	15
Piemonte	7	9	Puglia	17	18
Marche	8	12	Campania	18	20
Toscana	9	7	Sicilia	19	19
Liguria	10	8	Calabria	20	17

References

1. Bocci, L., Vicari, D.: Analysis of three-way data: an extension of the STATIS method. Book of Short Papers SIS 2021, pp. 627-632. Pearson (2021).
2. Burchi, F., Gnesi, C.: A review of the literature on well-being in Italy: A human development perspective. Forum for Social Economics **45**, 170-192 (2016).
3. Escoufier, Y.: L'analyse conjointe de plusieurs matrices de données. In : Jolivet, M. (eds.) Biométrie et Temps, pp. 59-76. Société Française de Biométrie, Paris (1980).
4. Istat: BES 2020 Report: Equitable and sustainable well-being in Italy. Istituto nazionale di statistica, Roma (2021).
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. Stiglitz, J. E., Sen, A., Fitoussi, J.-P.: Report by the commission on the measurement of economic performance and social progress (2009).

Multiway approach for clustering time series with time varying parameters

Approccio multiway per clustering di serie storiche con parametri time varying

Roy Cerqueti, Raffaele Mattera and Germana Scepi

Abstract In this paper, we propose a new approach for clustering time series with time varying parameters. Because of the time array nature of the dataset, we adopt a multiway approach. For showing the validity of the clustering algorithm a simulation study is produced.
Abstract *In questo articolo si propone un approccio di clustering di serie storiche con parametri che cambiano nel tempo. Data la natura del dataset considerato, si adotta un approccio per dati multiway. Per dimostrare la validità dell'algoritmo di clustering viene utilizzato uno studio di simulazione.*

Key words: Generalized Autoregressive Score (GAS), time varying parameters, Time series clustering, Spectral density, Multiway data analysis

1 Introduction

Clustering is one of the most important data mining algorithm, usually implemented for exploratory purposes, but also for more complex tasks like anomaly detection or classification. The idea of clustering time series on the basis of their distribution characteristics is mainly due to [12] that considered both skewness and kurtosis. Successively, [13] and [10] proposed approaches of clustering based on multiple features including static mean, variance, skewness and kurtosis. [7] proposed to cluster

Roy Cerqueti
Department of Economics and Social Sciences, Sapienza University of Rome e-mail: roy.cerqueti@uniroma1.it

Raffaele Mattera
Department of Economics and Statistics, University of Naples "Federico II" e-mail: raffaele.mattera@unina.it

Germana Scepi
Department of Economics and Statistics, University of Naples "Federico II" e-mail: scepi@unina.it

time series using extremes, i.e. according to static parameters estimated from a Generalized Extreme Value (GEV) distribution. In a similar fashion, [11] considered an approach of clustering based on parameters estimated by a Skewed Generalized Error Distribution (SGED).

The aforementioned approaches can be formalized as follows. Let $\mathbf{Y}\{y_{n,t} : n = 1,\ldots,N; t = 1,\ldots,T\}$ be the matrix containing the N time series of length T generated by a probability density function $p(\cdot)$ that is characterized by the presence of J parameters. The number of J parameters depends by the underlying distributional assumption. For example, if the distribution shows a Gaussian density, we have $J = 2$ because $p \sim N\left(\mu, \sigma^2\right)$, where μ is the mean and σ^2 is the variance. According to previous studies (e.g. see [7, 11]) it is possible to classify time series with similar mean μ and/or variance σ^2. Clearly, within this framework it is possible to consider alternative distributions that include more or less J parameters.

The drawback of these approaches lies on the fact that they consider static distribution features (i.e. static mean, static variance, etc.), while in time series context is very likely that these features are time varying. Despite clustering techniques based on time series' distribution characteristics have been extensively studied, an approach based on time varying parameters has been only recently explored by [2]. By considering a Gaussian density $p(\cdot)$ with time varying parameters, we have that $p \sim N\left(\mu_t, \sigma_t^2\right)$. [2] proposed to cluster time series according to a target time varying parameter, i.e. either μ_t or σ_t in the Gaussian case. Differently from [2], in this paper we propose a novel clustering approach that, instead of choosing a single target parameter, is based on multiple time varying parameters.

The consideration of more than one time varying parameters in the clustering process has a consequence in terms of the dataset structure. Indeed, for each n-th time series there are $J \geq 1$ parameters that vary over time as well. In other words, we have to deal with a *multiway* data structure [3]. In particular, we assume a time data array [6] where three dimensions (N time series, J parameters, T time) are considered.

The paper is structured as follows. In the next section, the proposed clustering approach is described. Then, in the section 3 a simulation study is produced. Final remarks are in section 4.

2 Clustering with time-varying parameters

The first step of the proposed clustering approach lies on the estimation of the time varying parameters. Let us suppose the most simple case, where the time series follow a Gaussian density $p \sim N\left(\mu_t, \sigma_t^2\right)$. Therefore, we have to estimate the time varying mean μ_t and the time varying variance σ_t^2.

In order to model and estimate the time varying parameters we use the Generalized Autoregressive Score (GAS) model [4]. The GAS model is based on the assumption that each n-th time series is generated by the following observation density $p(\cdot)$:

$$y_{n,t} \sim p(y_{n,t}|f_{n,t}, \mathscr{F}_{n,t}; \theta_n), \tag{1}$$

where θ_n is a vector of static parameters, $\mathscr{F}_{n,t}$ is the information set at time t, $f_{n,t}$ is a vector of length $J(j = 1,\ldots,J)$ of time-varying parameters depending by the probability distribution. The model's information set at a given point in time t, $\mathscr{F}_{n,t}$, is obtained by the previous realizations of the time series $y_{n,t}$ and the time-varying parameters $f_{n,t}$. The Generalized Autoregressive Score of order one, the $GAS(1,1)$, can be written as:

$$f_{n,j,t} = \omega_{n,j} + \mathbf{A}_{n,j,1} s_{n,j,t-1} + \mathbf{B}_{n,j,1} f_{n,j,t-1} \tag{2}$$

where $\omega_{n,j}$ is a real vector and $\mathbf{A}_{n,j,1}$ and $\mathbf{B}_{n,j,1}$ are diagonal matrices. All the scalar parameters $\omega_{n,j}, \mathbf{A}_{n,j,1}, \mathbf{B}_{n,j,1}$ are collected in the vector θ_n. Moreover, $s_{n,j,t}$ is the *scaled* score of the conditional density (1) in a time t with respect to a j-th parameter of the n-th time series. In other words, in the GAS model we suppose that the evolution of the time-varying parameter vector $f_{n,t}$ depends both by a vector $s_{n,t}$, proportional to the score of the density, and by an autoregressive component.

Another useful feature of the GAS model is that the vector θ_n is obtained by maximum likelihood estimator (for the details see [4]). Once the quantities $\omega_{n,j}$, $\mathbf{A}_{n,j,1}$ and $\mathbf{B}_{n,j,1}$ are estimated, the time varying parameters can be obtained by the in-sample predictions $\hat{f}_{n,j,t}$ [2]. Then, the estimated time varying parameters $\hat{f}_{n,j,t}$ are used as the input of the clustering procedure.

The goal of the proposed clustering approach is to assign time series with similar (in some sense, we will enter the details below) distribution's parameters in the same cluster [2]. Once the J time varying parameters are estimated, we consider the dataset $\mathbf{F}\{f_{n,j,t} : n = 1,\ldots,N; j = 1,\ldots,J; t = 1,\ldots,T\}$ as the input of the clustering algorithm. Following [6, 8], we define $\mathbf{F}$ as a *three-way data time array*, because both the n-th time series and the j-th parameter are time varying.

In order to cluster the objects contained into a three-way data time array we follow the two-step procedure of [6]. In the first step we compute, for each n-th time series, the dissimilarity matrix $\mathbf{D}_n = \left\{d_{n,j,j'} : n = 1,\ldots,N; j, j' = 1,\ldots,J; j \neq j'\right\}$ between each pairs of the J time varying parameters observed at T times. In this context, the dissimilarity represents the degree to which, for a given n-th time series, two time varying parameters j and $j\prime$ show a common pattern over the time. Thus, we obtain N distance matrices $\mathbf{D}_n$. In the second step, we classify the N time series on the basis of a diversity measure between each pair of distances $\mathbf{D}_n$. Because the dissimilarity matrices are squared and symmetric with a null diagonal, we can vectorize their lower triangular $\mathbf{L}_n$ obtaining $\mathrm{vec}(\mathbf{L}_n)$. Then, we define the following pairwise Euclidean dissimilarity between the time series n and $n\prime$:

$$d_{n,n\prime} = \|\mathrm{vec}(\mathbf{L}_n) - \mathrm{vec}(\mathbf{L}_{n\prime})\| \tag{3}$$

As noted by [6], this approach is directly related with the Relationship Matrices Analysis [3]. Starting from (3), we apply the Partition Around Medoids (PAM) algorithm to obtain the clusters. The PAM algorithm leads to more interpretable results and it more robust to outliers than the k-means algorithm. Moreover, it is also faster in terms of computational time than the k-means.

In the end, to address the issue of number of clusters selection, we suggest to use the Average Silhouette Width (ASW) criterion, which is well established in literature [1].

3 Simulation study

To show the validity of the proposed clustering procedure, we provide an application to simulated data. The simulation scheme works as follows. We simulate $N = 5$ time series that are normally distributed with a time-varying mean $\mu_{1,t}$ and variance $\sigma^2_{1,t}$ whose process is is given by the GAS with parameters:

$$\omega_1 = (0.0490, 0.0154); \quad \mathbf{A}_1 = \begin{pmatrix} 0.0001 & 0 \\ 0 & 0.0534 \end{pmatrix}; \quad \mathbf{B}_1 = \begin{pmatrix} 0.0485 & 0 \\ 0 & 0.9891 \end{pmatrix}$$

Then, we simulate another set of $N = 5$ Gaussian time series with $\mu_{2,t}$ and $\sigma^2_{2,t}$ generated by a GAS process with the following parameters:

$$\omega_2 = (0.0840, 0.0456); \quad \mathbf{A}_2 = \begin{pmatrix} 0.00001 & 0 \\ 0 & 0.0139 \end{pmatrix}; \quad \mathbf{B}_2 = \begin{pmatrix} 0.0660 & 0 \\ 0 & 0.0968 \end{pmatrix}$$

We consider four different scenarios in terms of time series' length, namely $T = \{2000, 1000, 500, 250\}$. In the first two scenarios we consider long time series, while in the last two short ones.

By exploiting the fact that the time series are generated by a Gaussian distribution with different time varying-parameters, we consider a Gaussian-GAS for clustering. Clustering according to time-varying parameters of a Gaussian distribution is equivalent to clustering approach based on conditional moments [2].

Following [2] we use the ACF-based Euclidean distance [9] in computing $d_{n,j,j\prime}$. Then, the final distance between two time series n and $n\prime$, $d_{n,n\prime}$, is computed according the (3). The proposed clustering approach is compared with two alternative clustering models. The first one is represented by the crisp version of [2], where a target parameter is selected for clustering and the auto-correlation based distance is employed. The second clustering approach is based on the auto-correlation distance for the original time series, i.e. we consider the crisp version of [9]. In all the cases

we assume $C = 2$ clusters. The comparison is made in terms of the average Adjusted Rand Index (ARI) over 300 trials as in [5]. The results are shown in Tab. 1.

First of all, we notice that the proposed approach provides the best classification for all the considered simulated scenarios. Moreover, the clustering accuracy improves with increasing time series length. Indeed, we have that with short time series $T = 500$ the ARI is equal to 0.38, while with $T = 2000$ it takes value of 0.88.

The validity of time-varying parameters based clustering is highlighted also by the fact that the approach with targeting of [2] is very competitive with respect to the standard ACF-based clustering on the original time series. Furthermore, clustering based on variance leads to much more accurate results than the mean-based clustering, hence confirming the results of [2]. This evidence suggests that the proposed approach would benefit from the relevance of the mean in the clustering process. Indeed, in this simulation study, the mean-based clustering leads to inaccurate results while the time-varying variance is a very good target for clustering. However, if both time-varying parameters had provided good clustering results, the fact of considering both of them instead of just one target would have further improved the classification quality.

Table 1 Clustering results: average Adjusted Rand Index

Clustering approach	$T = 2000$	$T = 1000$	$T = 500$	$T = 250$
Proposed approach	**0.8828**	**0.6052**	**0.3892**	**0.1328**
Target parameter (μ_t) [2]	-0.0053	0.0110	-0.0055	0.0041
Target parameter (σ_t^2) [2]	0.8567	0.5529	0.3202	0.1024
ACF-based [9]	0.0082	0.0015	0.0002	0.0269

Note: Table reports the average Adjusted Rand Index (ARI) over 300 trials. An ARI value close to 0 indicates randomness in the partition, while a value close to 1 indicates a very good classification. The best approach is highlighted with the bold font.

4 Final remarks

In this paper, we showed how to cluster time series according to their time varying parameters. Differently from [2] we considered many parameters together in the clustering process instead of just one at time. To this aim, we adopted a multiway clustering approach. In the end, a simulation study showing the validity of the proposed clustering approach has been provided. Moreover, similarly to [14], we also developed an application to the clustering of services' performance. However, it is not reported here for the sake of brevity.

References

1. Batool, F., Hennig, C.: Clustering with the average silhouette width. Computational Statistics & Data Analysis **158**, 107,190 (2021)
2. Cerqueti, R., Giacalone, M., Mattera, R.: Model-based fuzzy time series clustering of conditional higher moments. International Journal of Approximate Reasoning **134**, 34–52 (2021)
3. Coppi, R.: An introduction to multiway data and their analysis. Computational statistics & data analysis **18**(1), 3–13 (1994)
4. Creal, D., Koopman, S.J., Lucas, A.: Generalized autoregressive score models with applications. Journal of Applied Econometrics **28**(5), 777–795 (2013)
5. Díaz, S.P., Vilar, J.A.: Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. Journal of classification **27**(3), 333–362 (2010)
6. D'Urso, P.: Fuzzy c-means clustering models for multivariate time-varying data: different approaches. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **12**(03), 287–326 (2004)
7. D'Urso, P., Maharaj, E.A., Alonso, A.M.: Fuzzy clustering of time series using extremes. Fuzzy Sets and Systems **318**, 56–79 (2017)
8. D'Urso, P., De Giovanni, L., Disegna, M., Massari, R.: Fuzzy clustering with spatial–temporal information. Spatial Statistics **30**, 71–102 (2019)
9. D'Urso, P., Maharaj, E.A.: Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets and Systems **160**(24), 3565–3589 (2009)
10. Fulcher, B.D., Jones, N.S.: Highly comparative feature-based time-series classification. IEEE Transactions on Knowledge and Data Engineering **26**(12), 3026–3037 (2014)
11. Mattera, R., Giacalone, M., Gibert, K.: Distribution-based entropy weighting clustering of skewed and heavy tailed time series. Symmetry **13**(6), 959 (2021)
12. Nanopoulos, A., Alcock, R., Manolopoulos, Y.: Feature-based classification of time-series data. International Journal of Computer Research **10**(3), 49–61 (2001)
13. Wang, X., Smith, K., Hyndman, R.: Characteristic-based clustering for time series data. Data mining and knowledge Discovery **13**(3), 335–364 (2006)
14. Yahyaoui, H., Own, H.S.: Unsupervised clustering of service performance behaviors. Information Sciences **422**, 558–571 (2018)

Session of solicited contributes SS12 – *Assessment of Management Quality*
Organizer and Chair: Clelia Fiondella

Using partial triadic analysis for depicting the temporal evolution of Italian private healthcare organizations

L' utilizzo dell'analisi triadica parziale per rappresentare l'evoluzione temporale degli ospedali privati italiani

Alessandra Belfiore, Corrado Cuccurullo and Massimo Aria

Abstract The economic-financial configuration of Italian private hospitals is little explored, together with the governance and strategic structures. The aim of this work is to bridge this with the help of a statistical-quantitative method, analyzing the evolution of the configurations of Italian private hospital. Methodologically, after having collected the balance sheet data of all the Italian RSAs and having built a dataset with the data of 2008, 2012 and 2016, a factor analysis (PTA) was combined with clustering techniques to trace the profile of the Italian Private Hospitals in their evolution from the beginning of the Return Plans at the end of their third three-year period. The work concludes with a methodological proposal for the use of statistical analyzes useful for the domain of accounting and budget data.

Abstract *La configurazione economico-finanziaria degli ospedali privati italiani è poco esplorata, insieme agli assetti di governance e a quelli strategici. Lo scopo di questo lavoro è di colmare questo con l'aiuto di un metodo statistico-quantitativo, analizzando l'evoluzione delle configurazioni delle case di cura italiane. Metodologicamente, dopo aver raccolto i dati di bilancio di tutte le RSA italiane e aver costruito un dataset con i dati del 2008, 2012 e 2016, è stata svolta un'analisi fattoriale (PTA) abbinata a tecniche di clustering per tracciare il profilo degli ospedali privati italiani nella loro evoluzione dal principio dei Piani di Ritorno alla fine del loro terzo triennio. Il lavoro si conclude con una proposta metodologica per l'utilizzo di analisi statistiche utili per il dominio dei dati contabili e di bilancio.*

[1] Alessandra Belfiore, PhD candidate in *Entrepreneurship and Innovation* – University of Campania "Luigi Vanvitelli"- Corso Gran Priorato di Malta, 81043 Capua CE - e-mail: alessandra.belfiore@unicampania.it - phone: +39 0823 675187

♠ Corrado Cuccurullo, Full Professor of *Management and Economics* – University of Campania "Luigi Vanvitelli", Caserta, Italy – e-mail: corrado.cuccurullo@unicampania.it

Key words: Partial Triadic Analysis, clustering techniques, Private hospital, financial profile, economic performance

1 Introduction

In this case study we investigate the potential of partial triadic analysis (PTA), a special kind of multivariate analysis [1]. PTA, also called X-STATIS, is an extension of principal components analysis (PCA). PTA is meant to perform a statistical analysis of experiments when the same variables are measured on the same individuals at different points in time [2, 3, 4, 5, 6, 7, 8].

PTA is a technique based on a simplified approach of three modalities of factor analysis [9, 10]. It involves three essential steps.

Its first step is called the interstructure. It corresponds to a global representation and gives the "importance" of each table. During this step, a matrix of scalar products between tables is computed.

The second step, the compromise computation and analysis, is the main step of the method. The compromise table is computed as the weighted mean of all the tables of the series, using the components of the first eigenvector of the interstructure as weights. This table has the same dimensions and the same structure and meaning as the tables of the series. The compromise table exhibits the best summary properties of the initial tables. It is analyzed by a PCA, providing a picture of the structures common to all the tables and a simultaneous representation of individuals and variables.

The third step of a PTA is the analysis of the intrastructure. The rows and columns of all the tables of the sequence are projected on the factor map of the PCA of the compromise as additional elements. This step summarizes the variability of the series of tables around the common structure defined by the compromise.

This work aims to map the configurations of Italian private hospitals, through PTA combined with clustering techniques, to answer the following research questions.

- RQ1. What aspects best explain the variability between private hospitals?
- RQ2. What are the main configurations of private hospitals?
- RQ3. How have these configurations changed over time? (2008; 2012; 2016)

2 Measures and methods

The aim of this study is to show how the use of statistical analysis is useful for the domain of accounting and budget data. By using the PTA, combined with a

clustering technique, we map: (i) the different corporate governance characteristics of the Private hospital; (ii) the most relevant economic-financial indicators.

Description of the variables used to build our collection and description of the different methodology used.

2.1 Data collection

We queried through the Aida database by Bureau van Dijk to find the governance and financial data of all the Italian private hospitals that duly filed the financial statements in the years 2008-2012-2016. We extracted the following information.

•The economic performance, through (i) ROA, (ii) Ebitda margin, (iii) net income/sales;

•The financial profile is understood as (i) debt/equity ratio, (ii) financial leverage, and (iii) interest coverage ratio, (iv) primary liquidity ratio.

The query counted 1165 firms under Ateco code 86.10.10. We proceeded filtered:

•only firms that presented the following legal forms: s.p.a. and s.r.l.;

•only firms that do provide ordinary or day hospitalization services;

Our final dataset is composed of 198 companies. (2008 - 2012 - 2016). The data is arranged in a list of three tables, corresponding to the three periods. Our data frame is panel data, with each table corresponding to a date.

2.2 Data analysis

To answer the first research question and therefore bring out the most characterizing (and differentiating) characteristics of the Italian private hospital, was used PTA. For the analysis, was used ade4, a multivariate data analysis package for the R statistical environment [11]. We calculate the PTA of the economic-financial variables measured on 198 nursing homes but in three different periods. The three tables, which form our historical series, are standardized for each year and then transformed into a single data frame.

To answer the second research question, which is to identify the configuration of Italian private hospitals, was used agglomerative hierarchical clustering technique. Clustering allowed us to select and group private hospitals based on similarity measures [12,13]. The hclust function of the R stats package was used for the hierarchical analysis of the clusters.

Finally, to answer the third research question, that is to identify how the configuration of Italian private hospitals changes over time, we combined the results of PTA with the results of hierarchical analysis of the clusters. We identified the centroids of each cluster and analyzed their trajectory over time (2008-2012-2016).

We choose to use the centroids because studying the trajectory of all 198 private hospitals would lead to poorly understood results.

3 Findings

Results of PTA, results of agglomerative hierarchical clustering technique and results of the centroids analysis are shown below.

3.1 PTA

The first two components summarize 58% of the variance of the original variables. Since we are in a compromise situation, we believe that these two axes (Factor 1 and Factor 2) represent satisfactorily the information contained in the original variables. Figure 1 reports the coordinates of the Intrastructure step of the PTA and shows the factorial. The three-period graphs have the same scale and can be overlaid and compared. The advantage of using the PTA lies in the fact that all points are in the same space, so the two axes have the same meaning in all three graphs. We can interpret the first axis as the financial profile axis and the second axis as the profitability (economic performance). Through PTA, these two aspects have been identified as the best aspects to explain the variability among private hospitals.

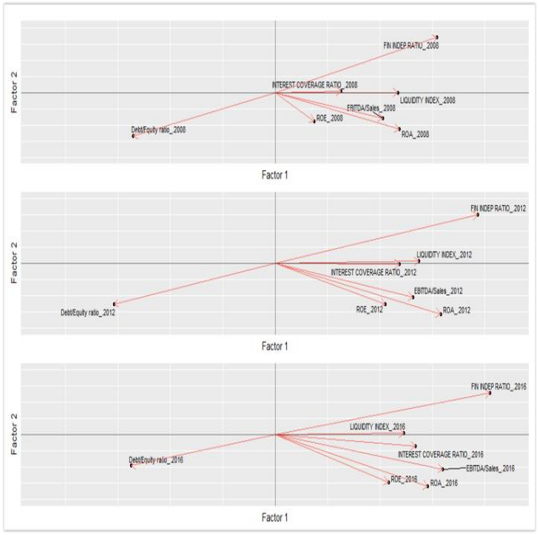


Figure 1: Factorial map of the economic-financial variables

3.2 Hierarchical clustering

Starting from the common structure of the axes over the years, it was then possible to identify the main configurations of private hospitals, through hierarchical cluster analysis. The results of the hierarchical analysis cluster, in the three periods, are presented in Figure 2. The 198 private hospitals form 4 different clusters. Looking at the average values of the variables in each cluster, we can understand what are the configurations of the private hospitals that compose them.

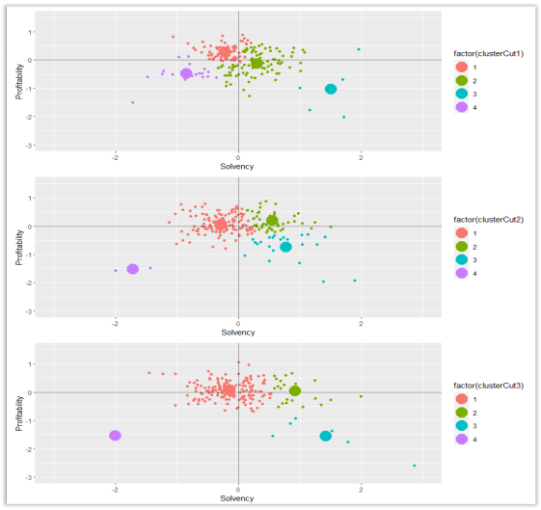


Figure 2: Factorial map of private hospital in 2008, 2012, 2016

Cluster 1 is the one located in the upper left quadrant. This cluster is characterized by private hospitals in a situation of severe financial stress. This cluster is defined by a negative liquidity index and a negative interest coverage rate. This means that this cluster has a shortage of liquidity with respect to short-term debts, and that the income generated is not sufficient to remunerate the capital acquired to produce it. In this case, the final solvency is negative.

Cluster 2 is the one located in the upper right quadrant. This cluster is characterized by private hospitals in an excellent financial and economic situation. In this cluster, we find positive values both of the debt ratios and of the profitability ratios. In this case, the solvency ability and profitability are positive.

Cluster 3 is the one located in the lower right quadrant. This cluster is characterized by private hospitals in a situation of severe economic tension. This cluster is defined by an EBITDA/sales ratio of less than 10% that means that you are not profitable companies. Also, ROE values of private hospitals in cluster 3 are low. This means that wealth is neither being created nor destroyed. In this case, the final profitability is negative.

Finally, cluster 4 is located in the lower-left quadrant. This cluster is made up of private hospitals in the worst situation. Indeed, in this cluster we find lower average values, both for debt ratios (eg liquidity index) and profitability ratios (ROE). In this case, the solvency and the final profitability are negative.

3.3 Combined analysis

Figure 3 shows graphic representation and trajectory of the centroids. The size of the spheres tells us whether that cluster has grown in number or not, while the arrow shows us how it has moved over time.
Cluster 1 increases in number and shifts down from factor 2 in 2012 and then shifts slightly to the right in 2016.Therefore, hospitals of cluster 1 already characterized by financial stress are also starting to have economic tensions.
Cluster 2 empties and rises along Factor 2, in 2012 and then moves decisively to the right, along Factor 1, in 2016. Therefore, hospitals of cluster 2, which we have said to be the best, decrease in number but continue to improve both financial and economic profiles.
Cluster 3 do not change in number but moves upwards in 2012 falling lower than before in 2016. The final result of the displacements of this cluster is a worsening of economic performance. Therefore, hospitals in cluster 3 that already had reduced profitability see their economic tension increase.
Finally, cluster 4 suddenly decreases moves decisively down, over time. This collapse is due to the worsening of both the economic and financial profile of the companies that formed this cluster. Therefore, hospitals in cluster 4, which we said are the worst, decreases but continue to deteriorate.

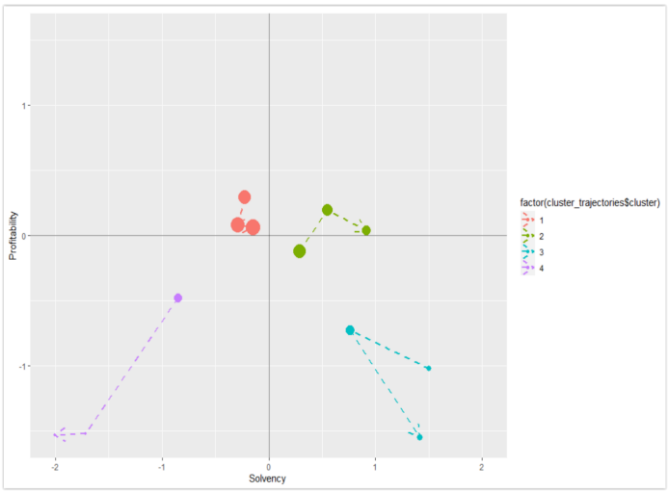


Figure 3: Graphic representation and trajectory of the centroids

4 Conclusions

This study has important implications for research. It must be said that the definition of the essential characteristics of private hospitals represents the first empirical challenge of the research that seeks to provide useful suggestions for practice. The need for contact between the academic and business world is very relevant especially in an area such as that of strategy and governance. Although configurational studies represent a promising development, they require appropriate research methodologies to realize their potential. In particular, since configurational research is based on the premise that there are distinct and meaningful groupings of companies within a larger entity, grouping methods are central to this avenue of investigation. This is why this study has great relevance in terms of originality. Not only for having considered a setting that is little considered as private hospitals but above all for having identified the best statistical method to analyze the data.

References

1. Jaffrenou, P.A.: Sur l'Analyse des Familles Finies de Variables Vectorielles: Bases Algebriques et Applications a la Description Statistique, These de Troisieme Cycle. Universite de Lyon, (1978)
2. Thioulouse, J., Chessel, D.: Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta oecologica. Série Oecologia generalis **8**(4), pp.463-480, (1987)
3. Simier, M., Blanc, L., Pellegrin, F., Nandris, D.: Approche simultanée de K couples de tableaux: application à l'étude des relations pathologie végétale-environnement. Revue de statistique appliquée **47**(1), pp.31-46, (1999)
4. Thioulouse, J., Simier, M., Chessel, D.: Simultaneous analysis of a sequence of paired ecological tables. Ecology **85**(1), pp.272-283, (2004)
5. Rolland, A., Bertrand, F., Maumy, M., Jacquet, S.: Assessing phytoplankton structure and spatio-temporal dynamics in a freshwater ecosystem using a powerful multiway statistical analysis. Water Research **43**(13), pp.3155-3168, (2009)
6. Bertrand, F., Maumy, M.: Using partial triadic analysis for depicting the temporal evolution of spatial structures: assessing phytoplankton structure and succession in a water reservoir. Case Studies In Business, Industry And Government Statistics **4**(1), pp.23-43, (2010)
7. Mendes, S., Gómez, J.F., Pereira, M.J., Azeiteiro, U.M., Galindo-Villardón, M.P.: The efficiency of the Partial Triadic Analysis method: an ecological application. Biometr Lett **47**, pp.83-106, (2010)
8. Thioulouse, J.: Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. The Annals of Applied Statistics **5**(4), pp.2300-2325, (2011)
9. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika, **31**(3), pp.279-311, (1966)
10. Kiers, H.A.: Hierarchical relations among three-way methods. Psychometrika **56**(3), pp.449-470, (1991)
11. Thioulouse, J., Dray, S.: Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. Journal of Statistical Software **22**(5), pp.1-14, (2007)
12. Romesburg, C.: Cluster analysis for researchers. Lulu.com, (2004)
13. Bridges Jr, C.C.: Hierarchical cluster analysis. Psychological reports **18**(3), pp.851-854, (1966)

Management of the human factor into the company. An experience from the aeronautic sector.

Gestione del fattore umano in azienda. Un'esperienza dal settore aeronautico.

Luigi Bollani, Alessandro Celegato, Filippo Barbero, Filippo Fontemaggi

Abstract This work starts from a longitudinal examination based on official statistics in the context of the flight and places a specific catastrophic event in the general context to determine its causes and opportunities for correction for the future. In particular, the mixture of technical problems with the management of the human factor is highlighted: the simultaneous exit of both controls makes the accident inevitable in the case examined. The experience that has marked the aeronautical world in this regard can be transferred to the corporate world to improve processes. In fact, the rigor that is necessary in flight management is more stringent due to the seriousness of possible errors, also due to situations of human stress, and the methods studied in aeronautics to deal with these problems can bring great benefit in the management of processes and personnel in agency.

Abstract *Questo lavoro parte da un'esame longitudinale basato sulle statistiche ufficiali nell'ambito del volo e situa un evento catastrofico specifico nel contesto generale per determinarne le cause e le opportunità di correzione per l'avvenire. Si evidenzia in particolare la commistione di problemi di carattere tecnico con la gestione del fattore umano: l'uscita di controllo contemporanea di entrambi rende, nel caso, inevitabile l'incidente. L'esperienza che a questo proposito ha segnato il mondo aeronautico può essere trasferita nel mondo aziendale per migliorare i*

Luigi Bollani, ESOMAS Department, University of Turin, Accademia Italiana del Sei Sigma (AISS); email: luigi.bollani@unito.it

Alessandro Celegato, Accademia Italiana del Sei Sigma (AISS), PSV Project Service and Value, email: alessandro.celegato@gmail.com

Filippo Barbero, F.B.F. Solution, email: filippo.barbero10@gmail.com

Filippo Fontemaggi, F.B.F. Solution, filippo.fontemaggi@libero.it

processi. Infatti il rigore che si rende necessario nella gestione del volo è più stringente per la gravità di possibili errori, dovuti anche a situazioni di stress umano e i metodi studiati in areonautica per fronteggiare questi problemi possono portare grande giovamento nella gestione dei processi e del personale in azienda.

Key words: human factor, improvement actions, aviation sector, fly accidents

1 Reason

This work stems from the human, research and professional experience of the authors and their observation of the business world, expressed above all through their affiliation with the Italian Academy of Six Sigma (AISS).
In the following, an account is given of how individually remediable risk or error circumstances may arise jointly, producing conditions that can get out of control and produce sometimes even catastrophic effects.
In these circumstances, the human factor often plays a decisive role due to its lower programmability and emotional capacity to adapt to unexpected situations. Hence the importance of investing in strengthening awareness and stress management in all company roles and above all for the most critical activities for the safety, control and development of the company.
For example, the aeronautics and air rescue sectors are discussed, where the management and compensation of errors plays a prominent role in contrasting situations of high gravity and the use of empowerment techniques typical of the most qualified flight personnel is proposed for consider their inclusion also in different fields.

2 Fly statistics in the last two centuries

Data about fly are available from the beginning of the last century because of the immediate interest introduced by the first fly of Write brothers in 1903.
Besides the number of flights and vehicles in different years (or periods), a large part of statistics is dedicated to various kind of accidents. On this subject it is important to share some unified vocabulary:
a) *accident* is an occurrence between the time any person boards the aircraft with the intention of flight until such time as all such persons have disembarked, in which a person is fatally or seriously injured and/or the aircraft sustains damage or structural failure engine and/or the aircraft is missing or is completely inaccessible.
b) *hijack* means the unlawful seizure or wrongful exercise or control of the aircraft (or the crew thereof).

c) *incident* is an occurrence, other than an accident, associated with the operation of an aircraft which affects or could affect the safety of operation.

d) *other occurrence*, meaning safety occurrences that cannot be defined as 'accident', or 'incident'. Usually aircrafts beyond repair for hurricanes, typhoons, sabotage and so on.

e) *unfiled occurrence*, when there is insufficient information to determine the exact type of occurrence.

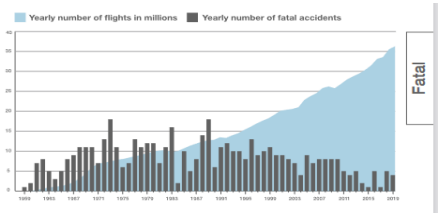
Of course there are many sources of data regarding flight; among the others EASA (European Union Aviation Safety Agency), Airbus (European producer), Allianz (Fly insurance), IATA (International Air Transport Association), ICAO (International Civil Aviation Organization), Boeing (Global producer).

To show some general trends, some data and graphics contained in the "Statistical Analysis of Commercial Aviation Accidents 1958-2019 - Airbus" are presented in the following (in fact the pandemic period can be considered as a local perturbation, not useful in this analysis).

The fatal accidents rate per million flights was recently 0.14 (2018) or 0.11 (2019); the flight departures were recently 35 million (2018) or 36 million (2019); the in-service fleet (aircrafts) was recently 25760 (2018) or 26680 (2019); historical data shows air traffic doubles about every 15 years.

In figure 1 yearly number of flights in millions and yearly number of fatal accidents are shown.

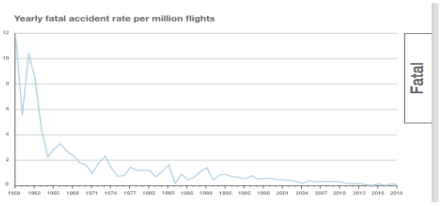
Figure 1: flights and fatal accidents (Source: Airbus)



The graph shows that the trend of accidents, not increasing along the whole period of time considered, is not driven by the trend of flights.

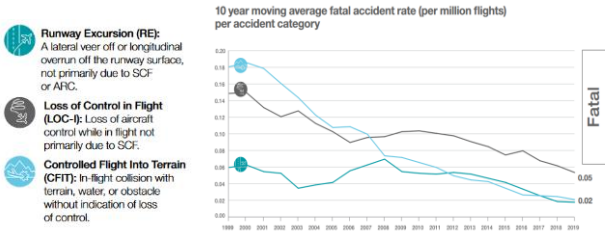
Consequently, the accident rate is decreased in the same period, as shown in figure 2.

Figure 2: fatal accidents rate (Source: Airbus)



The decreasing trend is greatly due to technical innovations.
Anyway different factors are to be considered important. Figure 4 shows that the accident rate due to Controlled Flight Into Terrain (CFIT) is heavily decreased in the last twenty years; the accident rate due to Runway Excursion (RE) is gone better starting from ten years ago and was not very high at the beginning of this century; the Loss of Control in Flight (LOC-I) presents recently the major rate of fatal accidents, even if twenty years ago it was nearly comparable to CFIT and much greater than RE. So LOC-I appears nowadays to be very difficult to contain and it is strictly linked to errors depending on human factor. In fact, there is a dependency of the pilot by the system and it is very difficult to consciously understand its indications.

Figure 3: fatal accidents rate (Source: Airbus)



3 A 2003 important accident

The year 2003 is an important year for the history of aviation for two specific events. The first since it was the centenary of the first flight. The second for a flight accident that occurred on January 8, 2003 in the city of Charlotte, North Carolina.
The Beechcraft 1900 airplane in service with Air Midwest while making Flight 5481 flew 37 seconds during take-off and then crashed near the US Airways hangar catching fire.
All 21 people on board were killed in the crash. The cause of the accident was identified in an incorrect attitude assumed by the aircraft, defined as aerodynamic stall.
The accurate investigations carried out by the National Transportation Safety Board showed that in fact the aerodynamic stall condition was determined by two factors.
The first was to be found in the calculation operations relating to the centering of the aircraft. The calculation made by the pilots was performed considering an average weight for each passenger equal to 80kg.
This value was in accordance with the provisions of the airline company and what defined by current regulations.

The technical investigations of the accident showed that in reality the average weight of the passengers was 90kg. This resulted in an overload of the aircraft which affected the positioning of the correct center of gravity.
The second factor that caused the accident was identified in incorrect maintenance performed on the aircraft itself. This had an influence on the aircraft's tail planes, limiting the efficiency of flight controls.
The Aircraft Accident Report highlighted that the accident occurred due to the simultaneous presence of the two factors, given that individually they were not sufficient to compromise the safety of the flight.

4 The human factor centrality

This specific incident, through investigations, has highlighted a single cause, "the human factor".
The cause of aircraft maintenance not performed in compliance with the requirements of the flight manual must be sought in the real context in which the airline's personnel were operating.
Both the operators and the management were subjected to particularly stressful conditions, leading them not to carry out the correct operations and furthermore the management itself was not aware that an Organized System was deviating from the operating procedures created precisely to guarantee flight safety.
Therefore, the methods of the human factor cannot be left to the management of the individual, but must belong to the know-how of the entire organizational structure at all levels, including control bodies.
The awareness that the average weight of the population (and therefore the passengers) has increased over time by 10kg, must have been a factor that the control bodies had to highlight.
But no aviation industry and no government body had understood that the numerous warning signs that came from the medical field, which emphasized that obesity was a major health emergency, could have an impact on flight safety.
If, on the other hand, the entire organizational structure of the flight had applied the human factor methods in a timely manner, it would have been possible to intervene in time, reaching the goal of "flying safely"

5 The contribution of aeronautic experience in companies

The path of the article highlights how the result of quality in service is conditioned not only by how the process itself is governed but also by how the human factor is managed.

In particular, in the aeronautical field, the quality level of flight safety is strongly linked to the ability to manage the human factor in critical situations; such as stress or, more generally, anything that can deviate from the conduct of the flight.
Therefore, this underlines the need for an effort to attribute an emotional competence to those who have to perform certain services.
The attention paid to highly specialized training in the aeronautical field can act as a guide for improvement in the training processes of other companies in other sectors, such as healthcare.
In this sense, the consolidated experience of AISS is reported, which aims to develop and disseminate company management tools.
AISS has found it necessary to combine the traditional methodology it proposes towards companies with the methods used in the aeronautical field.
Recently it is proposed to integrate among the tools of the Six Sigma which is composed of 5 steps (Define, Measure, Analyze, Improve and Control), with the human factor control method proposed by FBF Solution, in particular the "Flying Based" method (based on flight experience in the context of the Air and Rescue Frecce Tricolori).
This part of training in the human factor focuses on the following aspects: Training under stress, error management and stress control.

Conclusions

It has been highlighted that some defects in the control of regulatory evolution, in this case represented by a failure to take into account the demographic evolution of the population, can lead to catastrophic risk circumstances.
It is believed that this neglect is also to be found in other sectors not far from everyday life (e.g. does the design of a bridge designed for the traffic of 50 years ago need to be overhauled over time?).
In the specific circumstance of an impending risk, it was shown how the correct intervention of the human factor can be fundamental and how qualities of emotional maturity and stress control, which can be reinforced with adequate training processes, are fundamental in emergency management.
Finally, we present the example of AISS which recently included in the training proposal a human factor management module proposed by FBF and concerning emotional control.

References

1. Barbero, F., Fontemaggi, F., Celegato, A.: Human Factor and Error Management: Il Metodo Flying Based. Quality & Engineering **5**.**1**, 6--19 (2021)
2. International Civil Aviation Organization, Annex 13: Aircraft Accident and Incident Investigation (2020)
3. Airbus: Statistical Analysis of Commercial Aviation Accidents 1958-2019 (2020)

Session of solicited contributes SS13 – *New technologies for students learning assessment and evaluation*
Organizer and Chair: Alfonso Iodice D'Enza

A non parametric cognitive diagnostic method in classroom assessment conditions

Un modello diagnostico cognitivo non parametrico in condizioni di valutazione in classe

Evripidis Themelis, Angelos Markos

Abstract This paper presents a non-parametric approach to cognitive diagnostic assessment in classroom settings. The method improves upon the well-established general nonparametric classification (GNPC) method to obtain a fast and accurate classification of students' cognitive skills when the sample size is small. The proposed approach is based on the results of a simulation study using the Squared L_2 family of distances in the context of the GNPC method.

Abstract *Questo articolo presenta un approccio non parametrico alla valutazione diagnostica cognitiva in ambienti di classe. Il metodo migliora il consolidato metodo di classificazione generale non parametrico (GNPC) per ottenere una classificazione veloce e accurata delle abilità cognitive degli studenti quando la dimensione del campione è piccola. L'approccio proposto si basa sui risultati di uno studio di simulazione utilizzando la famiglia di distanze quadrate L_2 nel metodo GNPC.*

Key words: cognitive diagnosis, GNPC, Squared χ^2 distance

1 Introduction

The educational process is a dynamic and complex process. Through this process students are taught new knowledge and develop new skills. Cognitive Diagnostic Models (CDMs) are statistical methods used to categorize a group of students into homogeneous groups based on specific cognitive characteristics that they possess or not [12] .

Evripidis Themelis
Democritus University of Thrace, Alexandroupoli, e-mail: ethemeli@eled.duth.gr

Angelos Markos
Democritus University of Thrace, Alexandroupoli e-mail: amarkos@eled.duth.gr

CDMs provide detailed diagnostic feedback for whether a student has or has not master a number of skills based on a cognitive test. In contrast to the statistical methods used in the context of cumulative assessment with the main aim of classifying students based on their performance, CDMs are tools of formative assessment [1]. In other words, they do not quantify a student's level of knowledge with an overall score on a test (e.g., as the sum or the average of individual responses), but focus on the type of knowledge the student has or has not acquired. Through this process, the teacher can intervene with each student individually or intervene to different groups of students having a common profile of strengths or weaknesses.

In general, there are two types of CDMs in the relevant literature, parametric [6] and non-parametric CDMs [11]. Parametric models require a fairly large sample size (usually larger than 500), making it impossible to apply them to small student samples, such as those that are usually encountered in a classroom (usually smaller than 30). This gap is filled by non-parametric models, which, although they do not have the flexible probabilistic background of parametric models, can be effectively applied to when the sample size is small.

In the related literature, we identify four methods that fit into the non-parametric CDM framework, the Capability Matrix (CM) method [4], the Sum score Matrix (SsM) method [7], the non-parametric classification (NPC) method [8] and the general non-parametric classification (GNPC) method [9]. These exploratory or algorithmic methods are based on the calculation of a distance metric between the answer profile vector given by a student in a multiple choice test and the ideal answer profile that we would expect the student to give if he/she belonged to a certain skill or attribute profile. Simulation studies have shown that the four methods have better classification accuracy than the corresponding parametric CDMs when the sample size is relatively small [9], which makes them ideal for classroom-size applications. Among the four methods, the GNPC [9] was shown to be the most efficient in terms of classification accuracy.

2 The GNPC method

Let $\mathbf{Q}$ be a matrix a $J \times K$ matrix, where J is the number of dichotomous questions (i.e., correct/incorrect or 0/1) in a multiple-choice test and K is the number of skills. The elements of the Q matrix are 0 or 1, where $q_{ij} = 1$ if the i^{th} question requires the j^{th} skill and $q_{ij} = 0$ otherwise. The $\mathbf{Q}$ matrix is created by the test developer and needs to be properly structured (see [3] for the definition of $\mathbf{Q}$-matrix completeness). The general form of the Q matrix is:

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1k} \\ \vdots & \ddots & & \vdots \\ q_{i1} & q_{i2} & \cdots & q_{ik} \end{bmatrix}.$$

Let $\mathbf{Y}$ be a matrix of size $I \times J$, where I denotes the number of students/examinees. The elements of the $\mathbf{Y}$ matrix are 0 or 1, where $y_{ij} = 1$ if the i^{th} student answered the j^{th} question correctly and $y_{ij} = 0$ otherwise. It has the general form:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} \\ \vdots & \ddots & & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} \end{bmatrix}$$

The ideal profiles, $(a_1, a_2, ..., a_{2^K})$, express the ideal answers that students would give to belong to each profile. In their general form the ideal profiles of the students are given by the vectors:

$$a_1 = (0,0,...,0), a_2 = (1,0,...,0), a_3 = (0,1,...,0), \cdots, a_{2^k} = (1,1,...,1)$$

The GNPC [9] method starts by using the NPC [8] method in order to do the first classification of the student to an ideal profile a_m, choosing either DINA or DINO as the initial assessment model. After calculating the ideal response vectors from either

$$\eta_{ij}^c = \prod_{k=1}^{K} a_{ik}^{q_{jk}} \text{ (DINA) or } \eta_{ij}^d = 1 - \prod_{k=1}^{K} (1 - a_{ik})^{q_{jk}} \text{ (DINO)}$$

the distances between the ideal response profiles (η_{ij}^d or η_{ij}^c) and student responses (y_{ij}) are calculated with the use of the Manhattan distance [8].

This procedure i ideal response profiles. The method combines the two estimators from both the DINA and the DINO model used for the NPC method. In GNPC, the ideal answer profiles are calculated as follows:

$$\eta_{mj}^w = \eta_{mj}^d + w_{mj} \cdot (\eta_{mj}^c - \eta_{mj}^d) \tag{1}$$

The following cases can be distinguished [10]:

- 1^{st} case $\rightarrow \eta_{mj}^d = \eta_{mj}^c = 0 \Rightarrow \hat{\eta}_{mj}^w = 0$
- 2^{nd} case $\rightarrow \eta_{mj}^d = \eta_{mj}^c = 1 \Rightarrow \hat{\eta}_{mj}^w = 1$
- 3^{rd} case $\rightarrow \eta_{mj}^d = 1$ and $\eta_{mj}^c = 0 \Rightarrow \hat{\eta}_{mj}^w = \overline{Y}_{jm}$

where $\overline{Y}_{jm} = \dfrac{\sum_{i \in C_m} y_{ij}}{N_m}$ and N_m is the number of students belonging to the class of ideal profiles C_m.

Then, the method seeks to minimize iteratively the Euclidean Squared distance between the students' answers and the ideal answer profiles [9].

3 The Squared L_2 family of distances

The χ^2 distance family or Squared L_2 family of distances contains the following distance metrics: Euclidean Squared, Squared χ^2, Pearson's χ^2, Neyman's χ^2, Squared

χ^2, Probabilistic Symmetric χ^2, Divergence, Clark and the Additive Symmetric χ^2 [14]. It is trivial to show that all these metrics have the same minimum in the context of the GNPC and therefore the same $\hat{\eta}_{mj}^w$. That makes them comparable as they also satisfy the properties of stability, uniqueness and minimization (see [10] for a discussion of the statistical consistency of the GNPC). The Neyman χ^2 and the Additive Symmetric χ^2 distance, however, cannot be used into the GNPC framework, as they involve a division by 0.

To find the profile each student belongs to, the GNPC algorithm minimizes the distance between the ideal profile vector ($\hat{\eta}_{mj}^w$) and the answer vector of each student (y_{ij}). Minimizing a sum of J terms is the same as finding the minimum of each term and then get the sum of these J terms. Let $c_{mj} = \sum_{j=1}^{J} y_{mj}$, where $0 \leq c_{mj} \leq N_m$, so that

$$\hat{\eta}_{mj}^w = \frac{c_{mj}}{N_m} \tag{2}$$

Table 1 shows the term that is going to be added to the sum depending on the answer of the student.

Table 1 Comparison of the distances of χ^2 family - term to be added in the sum

student's answer	Squared χ^2	Euclidean Squared	Pearson χ^2	Probabilistic Symmetric χ^2	Divergence	Clark
$y_{ij}=1$	$\frac{(N_m-c_{mj})^2}{N_m(N_m+c_{mj})}$	$\frac{(N_m-c_{mj})^2}{N_m^2}$	$\frac{(N_m-c_{mj})^2}{c_{mj}N_m}$	$2\frac{(N_m-c_{mj})^2}{N_m(N_m+c_{mj})}$	$2\frac{(N_m-c_{mj})^2}{(N_m+c_{mj})^2}$	$\frac{N_m-c_{mj}}{N_m+c_{mj}}$
$y_{ij}=0$	$\frac{c_{mj}}{N_m}$	$\frac{c_{mj}^2}{N_m^2}$	$\frac{c_{mj}}{N_m}$	$2\frac{c_{mj}}{N_m}$	2	1

With some basic algebra it can be shown that the Squared χ^2 distance contributes the smallest term to the sum when the student's answer is correct, bringing him/her closer to the ideal profile and on the opposite condition, in which the student's answer is wrong, the Squared χ^2 contributes a larger term, to the sum, moving the student away of the ideal profile. This makes the Squared χ^2 distance a suitable alternative for the GNPC method.

4 Experiments on simulated data

A series of experiments have been conducted in order to confirm the theoretical results. Two sizes of Q where considered, 20×3 and 20×4. The correlation of the simulated Y matrix was set to 0.8 and data were simulated based on seven different parametric CDM models (GDINA, DINA, DINO, ACDM, LLM, RRUM, mixed model). Six distance metrics of the L_2 distance family were compared, while the

number of the students was set to 10, 15, 20, 25 and 30. In order to obtain the Model Level Classification Accuracy each scenario was repeated 100 times.

Table 2 Model Level Classification Accuracy Test of the χ^2 distance family

models	Squared χ^2	Euclidean Squared	Pearson χ^2	Probabilistic Symmetric χ^2	Divergence	Clark
$I = 25$						Q matrix 20×3
GDINA	0.684 (0.10)	0.684 (0.10)	0.475 (0.12)	0.686 (0.10)	0.422 (0.11)	0.416 (0.13)
DINO	0.843 (0.08)	0.842 (0.08)	0.522 (0.14)	0.845 (0.08)	0.490 (0.15)	0.485 (0.15)
DINA	0.795 (0.10)	0.834 (0.08)	0.590 (0.14)	0.795 (0.10)	0.411 (0.14)	0.422 (0.14)
ACDM	0.665 (0.11)	0.655 (0.10)	0.448 (0.12)	0.656 (0.11)	0.422 (0.12)	0.414 (0.12)
LLM	0.650 (0.10)	0.648 (0.11)	0.458 (0.13)	0.657 (0.11)	0.431 (0.13)	0.439 (0.13)
RRUM	0.787 (0.09)	0.793 (0.08)	0.546 (0.12)	0.783 (0.09)	0.465 (0.11)	0.470 (0.13)
ALL	0.678 (0.10)	0.666 (0.11)	0.510 (0.12)	0.675 (0.10)	0.567 (0.13)	0.554 (0.13)

[1] mean values (standard deviation values)

Table 3 Model Level Classification Accuracy Test of the χ^2 distance family

models	Squared χ^2	Euclidean Squared	Pearson χ^2	Probabilistic Symmetric χ^2	Divergence	Clark
$I = 20$						Q matrix 20×4
GDINA	0.478 (0.11)	0.464 (0.11)	0.365 (0.13)	0.475 (0.12)	0.384 (0.11)	0.377 (0.12)
DINO	0.655 (0.12)	0.653 (0.11)	0.511 (0.15)	0.644 (0.12)	0.429 (0.13)	0.446 (0.13)
DINA	0.598 (0.11)	0.619 (0.11)	0.480 (0.12)	0.597 (0.11)	0.347 (0.13)	0.345 (0.11)
ACDM	0.446 (0.12)	0.432 (0.13)	0.363 (0.11)	0.443 (0.12)	0.331 (0.11)	0.340 (0.11)
LLM	0.447 (0.13)	0.441 (0.12)	0.357 (0.11)	0.440 (0.13)	0.338 (0.11)	0.334 (0.11)
RRUM	0.522 (0.11)	0.510 (0.12)	0.413 (0.12)	0.535 (0.11)	0.353 (0.11)	0.363 (0.11)
ALL	0.485 (0.11)	0.464 (0.12)	0.363 (0.10)	0.480 (0.11)	0.454 (0.10)	0.463 (0.11)

[1] mean values (standard deviation values)

In almost every case the Squared χ^2 distance and the Probabilistic Symmetric χ^2 distance performed better than the Euclidean Squared distance. The use of both distance metrics improves the classification of students in ideal profiles by 2% to 3%. The difference between the Squared χ^2 distance and the Probabilistic Symmetric χ^2 is less than 0.5%.

5 Conclusions - Discussion

This work proposed a modified GNPC method that leads to a more accurate classification of students in skill profiles and can subsequently lead to safer conclusions about whether or not these students possess the skills and knowledge that they have

been taught in the classroom. Based on the study results we recommend the use of the Squared χ^2 distance (or the Probabilistic Symmetric χ^2 distance) instead of the Euclidean Squared distance.

References

1. Robitzsch, A. and George, A. C.: The R package CDM for diagnostic modelling. In: Handbook of Diagnostic Classification Models, pp. 549-572. Springer, Cham (2019)
2. MacQueen, J.: Some methods of classification and analysis of multivariate observations. In: Le Cam L.M. and Neyman J. (eds.) Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-287. University of California Press (1976)
3. Köhn, H.-F. and Chiu, C.-Y.: Conditions of completeness of the Q-matrix of tests for cognitive diagnosis. In: Van der Ark A. L., Bolt D. M., Wang W.-C., Douglas J. A. and Wiberg M. (eds.) Quantitative psychology research: The 80th annual meeting of the psychometric society, pp. 255-264. Springer (2015)
4. Ayers, E., Nugent, R. and Dean, N.: Skill set profile clustering based on student capability vectors computed from online tutoring data. In: EDM2008: 1st International Conference on Educational Data Mining.
http://eprints.gla.ac.uk/47662/ Cited 20-21 June 2008
5. Barnes, T.M.: The Q-matrix Method of Fault-tolerant Teaching in Knowledge Assessment and Data Mining. Ph.D. Dissertation, Department of Computer Science, North Carolina State University (2003)
6. Matthias von D. and Young-Sun L.: Handbook of Diagnostic Classification Models., Springer, Switzerland (2019)
7. Chiu, C.-Y., Douglas, J. A. and Li, X.: Cluster analysis for cognitive diagnosis: Theory and applications. Psychometrika **74**, 633–665 (2009)
8. Chiu, C.-Y. and Douglas, J. A.: A nonparametric approach to cognitive diagnosis by proximity to ideal response profiles. Journal of Classification **30**, 225–250 (2013)
9. Chiu, C.-Y., Sun, Y. and Bian, Y.: Cognitive diagnosis for small educational programs: The general nonparametric classification method. Psychometrika **83**, 355–375 (2018)
10. Chiu, C.-Y., and Köhn, H.-F.: Consistency theory for the general nonparametric classification method. Psychometrika **84**, 830–845 (2019)
11. Ma, C.,de la Torre, J. and Xu, G.: Bridging Parametric and Nonparametric Methods in Cognitive Diagnosis. arXiv preprint arXiv **15**, 409 (2006)
12. Rupp, A. A. and Templin, J. L.: Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-arty. Measurement **6**, 219–262 (2008)
13. Tatsuoka, K.: A Probabilistic Model for Diagnosing Misconceptions in the Pattern Classification Approach. Journal of Educational Statistics **12**, 55–73 (1985)
14. Sung-Hyuk C.: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International journal of Mathematical models and methods in applied sciences **4**, 300–307 (2007)

Hybrid unfolding models to Likert-scale data to assess distance learning perception in higher education

Modelli ibridi di sviluppo per dati su scala Likert per valutare la percezione dell'apprendimento a distanza nell'istruzione superiore

Iannario M., Iodice D'Enza A. and Romano R.

Abstract The Covid-19 pandemic forced students of any age and level of education to change the learning process, from learning in presence, to Distance Learning (DL). Such a relevant switch to DL has not been seamless for students, from both practical and psychological perspectives. In fact, students adaptation to DL process also depends on Covid-19 induced stress. Aim of the paper is to analyse an Italian university students survey on DL perception and Covid-19 related psychological effects, such as stress. The proposed approach implements a hybrid method that synthesizes the DL perception items into an ordinal response that is then regressed on the remaining items, to study the the effects on DL perception of further aspects (such as stress) and identify the most relevant covariates. The modeling phase consists of the implementation of the adjacent categories models to take into account the intensity of the opinions (students' feeling) at each end of the spectrum.

Abstract La pandemia di Covid-19 ha reso necessario il passaggio immediato alla didattica a distanza (DAD) per studenti di ogni grado. Il passaggio alla DAD ha avuto effetti pratici e psicologici sugli studenti. La percezione del passaggio alla DAD è stato anche influenzato dallo stress indotto dallo stato di pandemia. Il presente lavoro mira ad analizzare i dati relativi ad un'indagine sulla percezione della DAD da parte di studenti universitari, e gli effetti dovuti a fattori psicologici. L'approccio proposto prevede una prima fase in cui gli item che descrivono la percezione DAD vengono sintetizzati in una variabile ordinale. Nella seconda fase, gli item associati a fattori psicologici vengono regrediti sulla variabile di sintesi della percezione DAD ottenuta in precedenza.

Iannario M., Iodice D'Enza A.
Dipartimento di Scienze Politiche, Università degli Studi di Napoli Federico II
e-mail: maria.iannario@unina.it; iodicede@unina.it

Romano R.
Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Napoli Federico II
e-mail: rosaroma@unina.it

Key words: adjacent categories models, distance learning; joint data reduction

1 Introduction

During the peak of the Covid-19 pandemic Distance Learning (DL) became the only option to keep the education process going. The switch has been sudden and forced by the pandemic, therefore it was not seamless nor smooth. In fact, practical aspects aside, the level of adaptation of the students to the DL process is related to psychological aspects such as the stress for the fear of contagion and the social limitations. The impact of DL on students is multidimensional, and it requires multiple tools both in gathering information and in analysing them. In particular, we considered the scale proposed by [1] to study the DL perception in high education students; we also considered the 'student stress scale', proposed and validated by [16], as a measure of the potential psychological issues induced by the pandemic, with effects on the DL perception and the 'future career anxiety' scale [11] conceived to measure a unidimensional conceptualisation of anxiety. Therefore, a survey is considered that consists of items from the aforementioned scales, and it is structured in four item-blocks: the first block contains 19 items on students demographics and their proximity to Covid-19 cases; the second block is of 23 items that measure the DL perception of the students; the third block, with 7 items, measures students stress induced by Covid-19; the forth is composed by 5 items measuring anxiety for the future. The survey refers to 1592 students from 60 Italian Universities, with University of Naples and University of Bologna being the most represented, with a 25.9% and 18.5% share, respectively. The response option for the majority of items is a 4 levels Likert-type scale, ranging from *strongly disagree* to *strongly agree*.
Aim of the paper is to analyse the complex survey results via a hybrid method. In particular, the DL satisfaction related items, a domain of the DL psychometric scale consisting of six items, are synthesised into a meta-item, a single ordinal variable. The synthesis is obtained by a suitable joint data reduction (JDR) method. The meta-item is then regressed on the remaining items (stress-related, covid-related and demographics) by means of the adjacent categories logit models ([8]). The latter have been selected among the ordinal regression models because the interest is on the probabilities for adjacent categories, r and $r-1$, rather than different points in the cumulative distribution. The paper is structured as follows: Section 2 briefly describes the generation process of the ordinal response; Section 3 introduces the adjacent categories models while Section 4 illustrates the main results and concludes the paper.

2 JDR-based ordinal response

In order to synthesize the students satisfaction on DL assessment, we apply on the the DL-related items a joint data reduction approach. The definition of data reduction (DR) comprises different unsupervised learning approaches, that is data reduction (column-wise DR) and clustering (row-wise DR). Practitioners often apply column-wise DR before distance-based clustering, to mitigate the risk of the so-called curse of dimensionality: distances between any pair of points tend to converge in high dimensions, making it hard finding clusters. The two-step approach, often referred to as tandem approach, may fail since the dimension reduction step is independent from the clustering step, and it can be detrimental for the identification of clusters. In joint DR (JDR) approaches, the two steps are part of an iterative procedure, that alternately optimize one step, given the other. Different JDR methods have been proposed, for continuous ([3],[15]), categorical [9] and mixed-type variables (see [13] for a review). In this paper we refer to cluster correspondence analysis (cluster CA, [14]), a JDR method suitable for survey data. Data from multiple items are collected in the block matrix $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$, each block $\mathbf{Z}_j$ being the one-hot encoding of the item j, with $j = 1, 2, \ldots, p$. The application of cluster CA on the DL-related item leads to the definition of a cluster membership variable, and the cluster CA objective is

$$\min \phi_{\text{CCA}}(\mathbf{B}^*, \mathbf{Z}_K) = \left\| \mathbf{D}_z^{-1/2}\mathbf{M}\mathbf{Z} - \mathbf{Z}_K\mathbf{G}\mathbf{B}^{*\prime} \right\|^2 \quad \text{s.t.} \quad \mathbf{B}^{*\prime}\mathbf{B}^* = \mathbf{I}_d \tag{1}$$

where $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ is a centering operator, $\mathbf{B}^* = \frac{1}{\sqrt{np}}\mathbf{D}_z^{1/2}\mathbf{B}$, $\mathbf{D}_z = diag(\mathbf{Z}'\mathbf{Z})$, $\mathbf{B}$ is the item weights matrix, and $\mathbf{Z}_K$ is the indicator coding of the cluster membership categorical variable.
For the cluster CA application on DL item-blocks, the input parameter K is set to four, just like the levels of the Likert scales. The groups characterization is used to order the cluster membership levels. Figure 1 shows the items from the DL-related block that characterize each group: in particular for each item $\mathbf{Z}_j$, $j = 1, \ldots, p$, the standardized residuals matrix of the table $\mathbf{Z}_K'\mathbf{Z}_j$ is computed and the values for each of the p_j levels in each of the K groups are reported in the plot. Large residuals (in absolute value) indicate high group characterization (positive or negative) of the corresponding item levels.

3 Adjacent categories models

The adjacent categories model has the basic form

$$P(Y \geq r | Y \in \{r-1, r\}, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}'\boldsymbol{\beta}), \quad r = 2, \ldots, K.$$

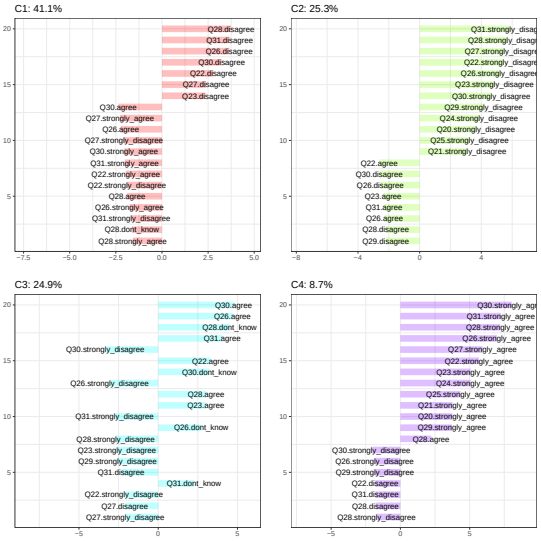


Fig. 1 Item scores for groups characterization: deviations from independence condition

Since $P(Y \geq r|Y \in \{r-1,r\},\boldsymbol{x}) = P(Y = r|Y \in \{r-1,r\},\boldsymbol{x})$ it specifies the probability of observing category r given the response is in categories $\{r-1,r\}$.

The most widespread model implies the use of the logistic distribution function for $F(.)$ yielding the (locally) logits for adjacent categories

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=r-1|\boldsymbol{x})}\right) = \beta_{0r} + \boldsymbol{x}'\boldsymbol{\beta},\ r = 2,\ldots,K.$$

For the obtained adjacent categories logit models the probabilities are

$$P(Y=r|\boldsymbol{x}) = \frac{\exp(\sum_{l=2}^{r}\{\beta_{0l} + \boldsymbol{x}'\boldsymbol{\beta}\})}{\sum_{s=1}^{k}\exp(\sum_{l=2}^{s}\{\beta_{0l} + \boldsymbol{x}'\boldsymbol{\beta}\})},\ r = 1,2,\ldots,K.$$

The model can also be seen as a submodel of the nominal multinomial logit model or the corresponding regression model obtained from the row-column (RC) association model considered by [6, 7] and [10].

For the analysed models we consider the constrained version (for relaxing this hypothesis see [5]) that is the effect β of each explanatory variable x on the odds of making the lower instead of the higher response is identical for each pair of adjacent response categories providing advantages of parsimony, such as a simpler summary of the effects. Furthermore results depend on the distance between categories, so this model uses the ordering of the response scale. For testing the proportionality assumption in our model we implement the Wald test proposed in [4].

4 Results

In studying DL students' satisfaction we favoured a regression method for ordinal outcomes that allows for comparisons of discrete, ordered, categories (ranging from *Not at all Satisfied* to *Very Satisfied*). The adjacent approach focuses on these local (adjacent) comparisons highlighting the intensity of opinions at each end of the spectrum; a result which is not possible in the standard cumulative odds model [12, p.215].

We control for several independent variables in the models based on previous studies of DL (see [2] and reference therein). Among covariates we found relevant *age* of students, *off site* referred to type of students who are not in site or commute, *negative* about students who have never been resulted positive to test for identification and isolation of infected persons and *no work*, an index that identifies full time students (no worker). Other items concerning student anxiety [11] and stress [16] scales resulted to be relevant: more specifically, from the former scale the item 'worry about future employment because the salary would probably not be as excellent as they wish for the devastating effect of Covid-19', from the latter scale the items concerning 'risk of contagion', 'social isolation' and 'academic studying experience'.

The estimates of the adjacent-categories model are in Table 1 with estimated thresholds $\hat{\beta}_{02} = -2.384$, $\hat{\beta}_{03} = -0.883$, $\hat{\beta}_{04} = 0.333$.

Table 1 Estimates for the adjacent categories model

	Coefficients	St.Err	t-stat
age	0.050	0.012	4.340
off site	-0.488	0.076	-6.430
negative	-0.211	0.068	-3.100
no work	0.224	0.073	3.070
anxiety	0.073	0.036	1.990
risk of contagion	0.162	0.037	4.410
social isolation	-0.244	0.038	-6.450
studying experience	-0.595	0.037	-15.860
Residual Deviance:	3442.421		
Log-likelihood:	-1721.211		
AIC:	3464.421		

Summarising results we observe that positive values of $\hat{\beta}_{age}$ and $\hat{\beta}_{nowork}$ imply that when *age* increases or we consider a full time student the probability of category r increases with respect to that of category $r-1$; that is respondent perceives a higher level of DL satisfaction. On the contrary the negative values $\hat{\beta}_{offsite}$, $\hat{\beta}_{negative}$, $\hat{\beta}_{social\,isolation}$ and $\hat{\beta}_{stud.experience}$ imply a decreasing DL satisfaction for increasing level of stress concerning isolation and academic studying experience lived during the pandemic period but also for students who no experienced the Covid-19 and are not in site. Positive values for the coefficients related to *risk of contagion* and

anxiety imply a higher evaluation in DL for respondents who perceive stressful the social relationships and generally perceive worry for the future.

Further developments may concern the replication of the survey to explore students' satisfaction for blended learning, consisting in the combination of distance learning with classroom learning and evaluate a hierarchical extension in which scrutinize the complex relationships among the meta item and covariates on different levels (universities).

References

1. Amir, L.R., Tanti, I., Maharani, D.A., Wimardhani, Y.S., Julia, V., Sulijaya, B., Puspitawati, R.: Student perspective of classroom and distance learning during covid-19 pandemic in the undergraduate dental study program universitas indonesia. BMC medical education **20**(1), 1–8 (2020)
2. Bacci, S., Fabbricatore, R., M, I.: Multilevel irt models for the analysis of satisfaction for distance learning during the covid-19 pandemic. Manuscript (2021)
3. De Soete, G., Carroll, J.D.: K-means clustering in a low-dimensional euclidean space. In: New approaches in classification and data analysis, pp. 212–219. Springer (1994)
4. Dolgun, S., Saracbasi, O.: Assessing proportionality assumption in the adjacent category logistic regression model. Statistics and Its Interface pp. 275–295 (2014)
5. Fullerton, A., Xu, J.: Constrained and unconstrained partial adjacent category logit models for ordinal response variables. Sociological Methods & Research **47**(2), 169–206 (2018)
6. Goodman, L.A.: Association models and canonical correlation in the analysis of cross-classification having ordered categories. Journal of the American Statistical Association **76**, 320–334 (1981a)
7. Goodman, L.A.: Association models and the bivariate normal for contingency tables with ordered categories. Biometrika **68**, 347–355 (1981b)
8. Goodman, L.A.: The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. Biometrics **39**, 149–160 (1983)
9. Hwang, H., Dillon, W.R., Takane, Y.: An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. Psychometrika **71**(1), 161–171 (2006)
10. Kateri, M.: Contingency table analysis. Methods and implementation using R
11. Mahmud M.S. Talukder M.U., R.S.: Does 'fear of covid-19' trigger future career anxiety? an empirical investigation considering depression from covid-19 as a mediator. The International journal of social psychiatry (2020)
12. Sobel, M.: Modeling symmetry, asymmetry, and change in ordered scales with midpoints using adjacent category logit models for discrete data. Sociological Methods & Research **26**, 213–232 (1997)
13. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics **11**(3), e1456 (2019)
14. van de Velden, M., Iodice D'Enza, A., Palumbo, F.: Cluster correspondence analysis. Psychometrika **82**(1), 158–185 (2017)
15. Vichi, M., Kiers, H.A.: Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis **37**(1), 49–64 (2001)
16. Zurlo, M.C., Cattaneo Della Volta, M.F., Vallone, F.: Covid-19 student stress questionnaire: Development and validation of a questionnaire to evaluate students' stressors related to the coronavirus pandemic lockdown. Frontiers in Psychology **11**, 2892 (2020)

Classification of Statistics learners using multi-dimensional latent class IRT model and archetypal analysis: the ALEAS app

Classificazione degli studenti di statistica con l'utilizzo di un modello IRT multidimensionale a classi latenti e l'analisi archetipale: l'app ALEAS

Daniela Pacella, Rosa Fabbricatore, Carla Galluccio and Francesco Palumbo

Abstract In recent years the need for assistive technologies for teaching and assessing Statistics is steadily rising. The ALEAS app aims to provide an intelligent adaptive assessment proposing a learning analytics approach that combines a psychometric model (multidimensional IRT) with Archetypal analysis to estimate the learners' ability in comparison with their peers.

Abstract *Negli ultimi anni è emersa sempre più la necessità di tecnologie a sussidio dell'apprendimento e insegnamento della statistica. L'app ALEAS propone una modalità di valutazione adattiva e intelligente utilizzando un approccio che unisce un modello psicometrico (IRT multidimensionale) con l'analisi archetipale per stimare l'abilità dello studente rispetto ai suoi pari.*

Key words: IRT, statistics, archetypal analysis, learning analytics

1 Introduction

The present study aims to describe an integrated methodological framework for teaching and assessing statistical knowledge with a focus on university students enrolled in non-scientific degree programmes that is developed and built in the ALEAS

Daniela Pacella
Department of Public Health, University of Naples Federico II, e-mail: daniela.pacella@unina.it

Rosa Fabbricatore
Department of Social Science, University of Naples Federico II, e-mail: rosa.fabbricatore@unina.it

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

Francesco Palumbo
Department of Political Sciences, University of Naples Federico II, e-mail: francesco.palumbo@unina.it

Fig. 1 Screenshots of the user interface and learning content of the ALEAS app.

(Adaptive LEArning system for Statistics) App. The ALEAS app aims to provide an adaptive learning environment that allows students to assess their own knowledge in Statistics. The statistical learning domain in ALEAS is arranged into a hierarchical structure defined by Areas, Topics and Units.

- For each Statistics Topic, a multidimensional IRT user model estimates items' difficulty and discriminating power and categorizes each student into a latent class based on the three Dublin descriptors *Knowledge and understanding*, *Applying knowledge and understanding* and *Making judgments*;
- at Area level, students are clustered into homogeneous groups by the archetypal analysis on their own average ability levels according to the latent class IRT models. Each student is classified into one of four performance profiles [1].

2 Methodology

The ALEAS app is a free mobile web-based application. The learning material consists of animated 3D cartoons, short stories and dynamically compiled exercises (an example is shown in Figure 1); content is dynamically generated from a Latex template containing R code. Exercises are organised in sets of 15 and can require an open or a multiple choice answer. After completing each exercise, the student is provided with the correct solution and a conceptual explanation.

2.1 Multidimensional IRT model

At the Topic-level, the ALEAS system exploits a two-parameter logistic (2PL) formalization [2] of the multidimensional latent class IRT models simultaneously considering more students' ability dimensions (Dublin descriptors).

More formally, the vector $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_D)'$ of the D latent variables (herein $D = 3$ according to the Dublin descriptors) is assumed to follow a discrete distribution with $\xi_1, \xi_2, \ldots, \xi_k$ support points defining k latent classes. Thus, students belonging to the same latent class share the same profile according to the Dublin dimensions defining students' ability in Statistics. The prior probabilities of belonging to latent classes determine the class weights $\pi_1, \ldots, \pi_k$, where $\pi_c = P(\Theta = \xi_c)$ with $c = 1, \ldots, k$, $\sum_{c=1}^{k} \pi_c = 1$ and $\pi_c \geq 0$. Given $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$ the possible realization of Θ with $\xi_c = (\theta_{c1}, \theta_{c2}, \ldots, \theta_{cD})$, the probability of correct answer to a

dichotomously-scored item i (with $i = 1,\ldots,I$) can be formalized as follows:

$$g[P(X_i = 1|\Theta = \xi_c)] = \log\frac{P(X_i = 1|\Theta = \xi_c)}{P(X_i = 0|\Theta = \xi_c)} = a_i\left(\sum_{d=1}^{D}\delta_{id}\theta_{cd} - b_i\right), \quad (1)$$

where $g(\cdot)$ is the logit link function; X_i is the response at item i with realization $x_i \in [0;1]$; δ_{id} is a dummy variable equal to 1 if the item i measures the latent trait d. Moreover, according to the 2PL parametrization, only the item discrimination a_i and the item difficulty b_i affects response probability.

Due to the assumption of *local independence*, the manifest distribution of the entire response vector $X = (X_1,\ldots,X_I)'$ can be expressed as:

$$P(X = x) = \sum_{c=1}^{k} P(X = x|\Theta = \xi_c)\pi_c, \quad (2)$$

where

$$P(X = x|\Theta = \xi_c) = \prod_{d=1}^{D}\prod_{i\in I_d} P(X_i = x_i|\Theta_d = \theta_{cd}). \quad (3)$$

The estimation of the model parameters is performed through the Maximum Marginal Likelihood (MML) approach [3].

Once the model parameters are estimated, each student is assigned to the class corresponding to the highest posterior probabilities of belonging [5].

2.2 Archetypal analysis

At Area-level, Archetypal Analysis (AA) further classifies students as follows. AA is an unsupervised learning technique seeking to synthesize multivariate observations using a reduced number of special vectors, that is, the *archetypes*.

Let X be a $n \times p$ data matrix with observations on rows and attributes on columns, and let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, $k << n$, be a reduced set of convex combinations of the observations, such that $\mathbf{z}_j = \sum_{i=1}^{n} b_{ij}\mathbf{x}_i$, $i = 1,\ldots,n$ and $j = 1,\ldots,k$. The aim is to approximate each observation $\mathbf{x}_i$ via the convex combination $\sum_{j=1}^{k} a_{ji}\mathbf{z}_j = \sum_{j=1}^{k} a_{ji}(\sum_{i=1}^{n} b_{ij}\mathbf{x}_i)$. In algebraic notation:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X}^\mathsf{T} - \mathbf{ZA}\|^2 = \min_{\mathbf{A},\mathbf{B}} \|\mathbf{X}^\mathsf{T} - \mathbf{X}^\mathsf{T}\mathbf{BA}\|^2 \quad (4)$$

where $\mathbf{A}$ and $\mathbf{B}$ are respectively $k \times n$ and $n \times k$ column stochastic matrices, that is, both $\mathbf{A}$ and $\mathbf{B}$ have non negative elements, and each of their columns add up to 1.

The problem does not admit solution in closed form; therefore [4] proposed an iterative procedure to optimize the loss function in Formula 4 alternately with respect to $\mathbf{A}$ for fixed $\mathbf{B}$, and the other way around. The convergence of the procedure is guaranteed, yet multiple random starts are needed to reduce the risk of local optima.

2.3 Preliminary results

Preliminary results on simulated response patterns of one Area consisting of 3 Topics [5] assuming $k = 4$ latent classes and $k - 1$ archetypes showed that the model is able to satisfyingly discriminate the learners' ability.

Regarding the Topic-level classification, the results of Topic 1 show that Class 1 includes subjects with poor performance in all the three Dublin descriptors domains; Class 2 encompasses subjects with good performance in Knowledge, average performance in Application and poor performance in Judgment; Class 3 regards subjects with average performance in all the three dimensions; finally, Class 4 consists of the subjects with good performance in all Dublin descriptor dimensions. It is worth noting that the Judgment domain (the most complex domain) reports the lowest score for all the latent classes. At Area level, we found that the first archetype represents students with poor performance in all Dublin descriptors; the second identifies students with better performance in Knowledge; the third archetype corresponds to students with good performance in all the three dimensions. Finally, the barycenter refers to an average performance. Student feedback is thus developed and provided upon this categorization.

3 Conclusion

We presented the rationale, methodological framework and preliminary application of the ALEAS App. ALEAS aims to provide an intelligent and adaptive classification of non-STEM degrees undergraduate students' statistical abilities integrating a supervised psychometric approach (multidimensional 2PL IRT) and an unsupervised algorithm (Archetypal analysis) to estimate the learners' statistical ability.

References

1. Fabbricatore, R., Parola, A., Pepicelli, G., Palumbo, F.: A latent class approach for advising in learning statistics: Implementation in the aleas system. CEUR Workshop Proceedings, 2817 (2021)
2. Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord M. R. Novick (Eds.), Statistical theories of mental test scores, pp. 395–479. Addison-Wesley (1968)
3. Thissen, D.: Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika **47 (2)**, 175–186 (1982)
4. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics, 36 (4), 338–347. (1994).
5. Pacella, D., Fabbricatore, R., Iodice D'Enza, A., Galluccio, C., Palumbo, F.: Teaching STEM subjects in non-STEM degrees: an Adaptive Learning model for teaching Statistics. in "Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology", Taylor Francis (*in press*).

Session of solicited contributes SS14 – *Labor Market and Enterprises*
Organizer and Chair: Lucio Masserini

Searching for new trends and dynamics in Labour Market: a statistical approach for the recruiting process

Nuovi trend e dinamiche nel mercato del lavoro: un approccio statistico per il processo di reclutamento

Gabriele Maggioni, Paolo Mariani, Andrea Marletta and Mariangela Zenga

Abstract In this paper, the roles of the most requested skills for getting a job in Italian Labour market are investigated using a new data set created by combining information from business and external sources. From a methodological point of view, a Conjoint analysis is performed to estimate the partial utilities for the most requested skills and their importance in defining the best combinations of skills to match job requirements. In particular, the profiles of workers recruited by The Adecco Group in Italy in the period 2016-2021 have been analysed detecting dynamics and movements during last years.

Abstract *Il contributo analizza il ruolo delle skills richieste nel mercato del lavoro italiano usando un nuovo dataset creato combinando dati provenienti da fonti aziendali ed esterne. Da un punto di vista metodologico, una Conjoint analysis è stata implementata ai dati per stimare le utilità parziali legate alle competenze più richieste definendone la migliore combinazione desiderata in un'ottica di reclutamento. In particolare, i profili dei lavoratori reclutati da The Adecco Group in Italia nel periodo 2016-2021 sono stati analizzati alla ricerca di trend e dinamiche negli ultimi anni.*

Key words: Italian labour market, Job matching, Skills analysis

Gabriele Maggioni
The Adecco Group, e-mail: gabriele.maggioni@adeccogroup.com

Paolo Mariani
University of Milano-Bicocca e-mail: paolo.mariani@unimib.it

Andrea Marletta
University of Milano-Bicocca e-mail: andrea.marletta@unimib.it

Mariangela Zenga
University of Milano-Bicocca e-mail: mariangela.zenga@unimib.it

1 Introduction

In social and economic systems, the role of labour is fundamental, both for the aspects strictly related to labour as a production factor and for the perspectives regarding workers. The access to the labour market represents a key point for the supply and demand side. About the supply, the role of knowledge, abilities and attitudes leads to the consideration of models and formative offers for their creation and implementation. On the other hand, about the demand, the economic context and the effect of technical progress activate examples of improvement in roles and difficulties in the definition of short-term scenarios.

According the World Economic Forum, more than half of all employees will request a re-qualification before 2022. Among these employees, one third will need further education for six more months, and one fifth will need further education for a longer period [6]. According to the International Labour Organization (ILO), enterprises and employers will need to make new investments to expand their involvement in the education, training and re-skilling of workers to support economic growth. Additionally, workers will need to pro-actively upgrade their skills or acquire new ones through training, education and learning to remain employable [2].

Competencies could assume a central role in the competitiveness of firms and workers; for this reason they could represent a keystone of the retribution. Competencies may become a candidate in the integration or substitution the remunerative parameters, thus serving as a new tool in the relationship between jobs and wages.

The analysis was based on research proposed by The Adecco Group in Italy on new hires starting from 2016 to 2021. Information regarding goodwill, albeit with a managerial and administrative slant, provides a source of knowledge structured on the basis of the criteria that companies adopt in their choices of workers who apply for job positions in their companies. The aim of this work is to understand whether it is possible to define a time trajectory for some professional roles detecting trends and dynamics useful in the recruiting process.

The paper is structured as follows: after the introduction, a second section is dedicated to the methodologies used to answer the research objectives. A third section will show the description of the dataset and some preliminary results.

2 Conjoint analysis and choice models

In this paper, a conjoint analysis has been applied for the study of the choice models [3, 5] of the companies, starting from the preference expressed by the companies with respect to different possible configurations of requirements related to the professional profiles. The value of the level of satisfaction obtained by a company with respect to the obtained requirements is designated as Utility.

The Utility function assigns a level of satisfaction to each requirement considered, in particular, in the form:

$$U = f(X) \tag{1}$$

where U is the utility level and X are the characteristics of the requirements.

The profile is determined by the assignment of a level to each requirement under examination; the number of the profiles depends on the number of the attributes and their categories. For the conjoint analysis, the preference and utility are in a bi-univocal correspondence: the more a candidate meets the requirements of a company, the more his/her use will lead to usefulness. The preference can be interpreted as the function of the levels of the characteristics of a candidate. Subsequently, based on the preferred choice of the company, partial utilities are calculated. They represent the importance associated with each level of the attributes and are called part-worth. Finally, the total utility is analysed as the sum or the product of partial utilities. From an analytical point of view, this modelling is expressed as follows:

$$U_j = \sum_{l=1}^{L} \sum_{k=1}^{K} u_{jkl} * x_{jkl} + e_j \tag{2}$$

where U_j the utility of the j-th profile, u_{jkl} the partial utility referred to the l-th level of the k-th attribute, x_{jkl} a dummy variable that assumes a value of 0 or 1 if the level l of the attribute k is absent or present in profile j and e_j is the random error [4]. For this case the choice of model [1] requires the construction of all candidate profiles a priori as a combination of all attributes and levels. Among these the only one that represented the choice of the company is the one related to the profile of the candidate launched.

3 Application

In this paper, the dataset is obtained as a merge of business sources in combination with external sources. Internal sources are represented by the Adecco Group database on job offers and necessary requirements for the hires. External sources are the ESCO (European Skills, Competences, Qualifications and Occupations) classification for abilities and skills for professional figures and the Italian National Collective Labour Agreement contracts.

Regarding the internal sources, two macro-categories of data were detected: Candidate and job offer. About candidate, data are present for registry information and previous work experience. On the other hand, about the job offer, the set of requested recruitments are represented for each position in terms of work experience, linguistic knowledge, etc. About the external sources, the database has integrated the following information through the ESCO database and Italian National Collective Labour Agreement contracts. The ESCO (European Skills, Competences, Qualifications and Occupations) Taxonomy is used as a dictionary, to describe, identify and classify professional figures, abilities and qualifications relevant to the European labour market. The second external source is about retributive tables provided for the national level of different contracts.

Since data are available for a period of six years, from 2016 to 2021 (provisional data until September 2021), the analysis could be repeated for each year in order to find differences in the selected period. Beyond the differences, it is possible to sketch a defined path over the entire period. This path could be represented from a graphical point of view through the use of a time trajectory. The statistical unit is represented by a person receiving a job, and there were more than 1.000.000 job positions divided into the following 9 industries: Production and Logistic, Food services, Commercial and Marketing, Human Resources, Legal and Finance, Medical and Pharmaceutics, Engineering, Tourism and Fashion, IT and Digital. In table 1, the distribution of the job positions over the entire period and the industries is displayed.

Table 1 Distribution of the job positions over the period and the industries, Italy, 2016-2021

Industry	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Production and Logistic	127.012	159.449	136.831	103.973	99.879	92.141
Food services	24.041	29.648	27.337	23.096	12.085	12.843
Commercial and Marketing	14.595	18.099	12.708	10.117	7.971	7.889
Human Resources	5.925	7.044	7.792	6.336	4.153	3.610
Legal and Finance	3.520	3.667	4.016	4.183	3.240	2.734
Medical and Pharmaceutics	2.508	2.369	2.275	1.958	2.074	1.310
Engineering	1.856	1.838	1.734	1.481	1.034	905
Tourism and Fashion	5.847	6.843	6.309	3.807	1.801	589
IT and Digital	727	705	751	685	497	458
Total	**186.031**	**229.662**	**199.753**	**155.636**	**132.734**	**122.479**

Source: elaboration on The AdeccoGroup data

As it is possible to note from the Table 1, some preliminary differences at industry level are present. If the sector with more job offers is Production and Logistic for the entire period, Tourism and Fashion had a clear decrease in last years passing from 3% in 2016 to 0.5% in 2021. Starting from these differences, it will be possible to detect changes also in terms of skill required for the recruitment.

References

1. Dagsvik, J.K. Random utility models for discrete choice behavior. An Introduction. Statistics Norway Research Department, Norway (1998).
2. International Labour Organization, Skills, knowledge and employability (2018).
3. Krantz, D.H. Conjoint measurement: The Luce-Tukey axiomatization and some extensions. Journal of Mathematical Psychology **2**, 248-277 (1964).
4. Luce, R.D., Krantz, D.H. Conditional Expected Utility. Econometrica **2**, 253-271 (1971).
5. Street, D.J., Burgess, L. The Construction of Optimal Stated Choice Experiments: Theory and Methods. Wiley, New York (2007).
6. World Economic Forum. The future of jobs report. World Economic Forum, Geneva, Switzerland, (2018).

A new definition of the professional figure Open Manager

Una nuova definizione della figura professionale dell'Open Manager

Paolo Bruttini, Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga

Abstract This study focuses on the manager's professional work and its evolution in the last years. In particular, the main focus is to detect the possible new dynamics in the managerial behaviour able to define the new professional figure of the 'open manager', based on some evidence derived from a survey conducted by interviewing a set of managers of Italian companies using a structured questionnaire. By using an agglomerative hierarchical cluster procedure with the Ward's method, results six different groups of managers with similar behaviours were defined, based on the responses to the questionnaire items.

Abstract *Questo studio si concentra sul lavoro professionale del manager e sulla sua evoluzione negli ultimi anni. In particolare, il focus principale è quello di rilevare le possibili nuove dinamiche nei comportamenti manageriali che possano definire la nuova figura professionale del "manager aperto", sulla base di alcune evidenze derivate da un'indagine condotta intervistando un insieme di manager di aziende italiane tramite un questionario strutturato. Utilizzando una procedura di clustering gerarchico agglomerativo con il metodo di Ward, sono stati definiti sei diversi gruppi di manager con comportamenti simili, sulla base delle risposte agli item del questionario.*

Key words: Open manager, Hierarchical cluster analysis, Ward's method

[1] Paolo Bruttini, Forma del Tempo; e-mail: pbruttini@formadeltempo.com
Paolo Mariani, University of Milano-Bicocca; e-mail: paolo.mariani@unimib.it
Andrea Marletta, University of Milano-Bicocca; e-mail: andrea.marletta@unimib.it
Lucio Masserini, University of Pisa; e-mail: lucio.masserini@unipi.it

1 Introduction

The labour market is a field in permanent evolution and this is also tangible in the realization of new professional roles or figures. Such evolution could be defined on the basis of the emergence of new tasks or alternatively it could be derived from a description of some behaviours that managers undertake in the course of their work. In this second case it is possible to talk about of evolution of a professional figure because even when managers perform the same tasks the behavioural approach could be different, for example in the field of interpersonal relationships or for other behaviours. In this context, this study focuses on the manager's professional work and its evolution in the last years. In particular, this aim of this paper is twofold: first, to detect the possible new dynamics in the managerial behaviour able to define the new professional figure of the 'open manager', based on some evidence derived from a survey conducted by interviewing a set of managers of Italian companies using a structured questionnaire; second, to validate the distributed questionnaire as a classification tool which could be useful for predicting the professional roles or figures of managers based on their reported managerial behaviours. The open manager figure is not actually well defined, so it appears to be as a latent figure, for this reason through the data analysis following this survey it was possible to outline some emerging attitudes and behaviours. Nonetheless, this concept was enhanced by some authors in combination with the definition of open innovation [1]. The paper is structured as follows: after the introduction, a second section is dedicated to data description and methodology, used to answer the research objectives whereas a third section shows some preliminary results.

2 Data and method

Data were collected by Fondirigenti and Confindustria in 2020 through a structured questionnaire distributed to two different sets of Italian companies and filled in by a managerial internal figure. The first group was composed by innovative companies, while the second one was obtained by selecting a set of generic firms. The total number of respondent in the two groups was equal to 383 managers, coming from 320 different companies. Of the 383 respondents, 213 belonged to the first group of firms and the remaining 170 to the second one. 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 socio-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. Such items were

formulated as a 5-point Likert scale, with responses ranging from 1 to 5 where 1 stands for "totally disagree" and 5 for "totally agree".

In order to identify homogeneous groups of managers who share common behaviors and attitudes among those described by the items of the questionnaire, an agglomerative hierarchical cluster analysis was carried out using the Ward's method [2, 3]. Agglomerative clustering works in a bottom-up manner that is, each observation is initially considered as a single-element cluster. Next, pairs of clusters are merged until all objects have been merged into a single cluster. In particular, Ward's minimum variance method minimizes the total within-cluster variance thus, at each step of the agglomerative procedure, the pair of clusters that leads to minimum increase in total within-cluster variance (or with the smallest between-cluster distance) are merged. To apply a recursive algorithm under this objective function, the initial distance between individual objects must be (proportional to) squared Euclidean distance. The result of hierarchical cluster analysis can be easily visualized using a tree-based representation of the objects, called dendrogram. Subsequently, the obtained groups were analysed in relationship with the single items of the questionnaire by comparing the answers distribution in each single group with that in the entire set of respondents. This allowed to locate some discerning items identifying the managers' behaviors useful for defining their openness level.

3 Results

Data analysis carried out using the agglomerative hierarchical cluster procedure with the Ward's method allowed us to define six different groups of managers with similar behaviours, based on the responses to the questionnaire items. Results are shown in Table 1, in terms of descriptive label of each group, their respective absolute number and the correspondent percentage. To choose the right number of clusters of the final solution shown below, the elbow method and the inspection of the dendrogram were considered.

Table 1: *Groups of managers defined after the hierarchical agglomerative cluster analysis using the Ward's method*

Group	***N***	***Percentage***
Group 1: Guardians traditionalist, defender	76	19.8
Group2 : Open Leaders	79	20.6
Group 3: Selfish people	69	18.0
Group 4: Regulators	67	17.5
Group 5: Explorers	73	19.1
Group 6: Opponents	19	5.0

The descriptive labels were assigned after identifying those items characterizing the behaviors of the managers in each group. The main features that distinguish the

managers in each group were derived by comparing the distribution of item responses for each group to the total, in order to bring out their level of openness. In shorts, the main characteristics of each group can be describe as follows.

- Group 1. Guardians traditionalist, defender: a) for the team to function, it is always necessary to clarify priorities; b) I feel that I am fond of my colleagues at this company; c) Sometimes I personally write the procedures that govern activities.
- Group 2. Open Leaders: a) I Don't like employees who can impose themselves on others; b) It is always prioritize the career development of your employees; c) Sometimes I do not personally write the procedures that govern activities.
- Group 3. Selfish people: a) Business today doesn't require the most consistency; b) I don't expect my employees to be able to make changes on their own; c) I feel that I am not fond of my colleagues at this company.
- Group 4. Regulators: a) I prefer collaborators who can impose themselves on others; b) Business today requires the utmost consistency; c) When faced with any critical task, I always know someone who can help me.
- Group 5. Explorers: a) I can accept constant change in the business world; b) It is important to admit your mistakes to co-workers; c) In the professional context, I act very quickly.
- Group 6. Opponents: a) I can't accept constant change in the business world b) It is not important to admit your mistakes to co-workers; c) I don't take every opportunity I get to learn new things.

References

1. da Mota Pedrosa, A., Välling, M., Boyd, B.: Knowledge related activities in open innovation: managers' characteristics and practices. International Journal of Technology Management **61**(3/4), 254-273 (2013).
2. Everitt, B.: Cluster Analysis. Heinemann Educational Books Ltd., London (1974).
3. Hartigan, J.A.: Clustering Algorithms. Wiley, New York (1975).

Session of solicited contributes SS15 – *Statistical Approaches to Environmental Sustainability*
Organizer and Chair: Alfonso Piscitelli

Measuring sustainability as an emergent property of whole system dynamics

Misurare la sostenibilità come una proprietà emergente della dinamica di sistema

Richard Aspinall

Abstract Agricultural land amounts to 37% of the global land area and 43% in Europe. It is estimated that more than 17% of the global direct emissions of greenhouse gases is due to agriculture, almost evenly accounted to crops and livestock. Therefore, pressure from the sector deserves attention in the light of global challenges. Modernization of agriculture in the last century determined by mechanization and constant innovations has enhanced productivity and boosted changes that raise fresh concerns about sustainability. The coupling of human and environmental systems adds a further layer of complexity to understanding. We suggest that only by apt statistical and non-classical approaches applied to complex dynamic systems and looking at changes in the perspective of system dynamics can enable capture of dampening effects that act to keep a dynamic equilibrium as well as address emergencies, i.e., new opportunities to regulate it.

Abstract *I terreni agricoli rappresentano il 37% della superficie terrestre globale ed il 43% in Europa. Si stima che oltre il 17% delle emissioni globali di gas serra sia dovuto direttamente all'agricoltura, quasi equamente ripartito tra le colture e l'allevamento. Chiaramente la pressione ambientale esercitata dal settore merita attenzione alla luce delle sfide globali. La modernizzazione dell'agricoltura nell'ultimo secolo, determinata dalla meccanizzazione e dalle continue innovazioni, ha aumentato la produttività e stimolato cambiamenti che sollevano nuove preoccupazioni sulla sostenibilità. L'interdipendenza tra i sistemi umano ed ecologico va ad aggiungere un ulteriore livello di complessità alla comprensione degli effetti di tali mutamenti. Suggeriamo che solo con opportuni approcci statistici applicati allo*

[1] Richard Aspinall, James Hutton Institute, Craigiebuckler, Aberdeen, Scotland, UK; email: rjaspinall10@gmail.com

studio dei sistemi dinamici complessi e con l'attenzione centrata sui cambiamenti, nella prospettiva dei sistemi dinamici, è possibile catturare gli effetti di smorzamento che agiscono per mantenere un equilibrio dinamico e cogliere le proprietà emergenti, tra le quali individuare anche opportunità che siano più sostenibili.

Key words: Land system, Coupled human-environment system, Sustainability accounting, Agriculture

1 Introduction

Land use, water and energy are recognised as exemplars of coupled human-environment systems [3]. Climate change, biodiversity loss, security of food, water and energy, human and environmental health, and questions of sustainability, are consequences of human actions within these coupled systems, yet addressing so many issues concurrently require improved understanding and tools for interpreting the complexity of interdependencies and functioning within whole systems [4]. As one example, consider agricultural land uses. Globally these occupy 37% of the total land area, and over half of habitable land [11]. The proportion is 43% of the total land area in Europe. The capacity for extension of agriculture into further land area is considered unlikely [8]. Additionally, about 17% of global greenhouse gases are derived from agricultural activities [6]. Agriculture thus not only provides food, but is also contributes to climate changes, itself threatening agricultural production, and is associated with losses of biodiversity, degradation of water quality, and with other impacts on a variety of ecosystem goods and services [7].

An aspiration and research goal of land system science is to develop understanding of land systems within the context of sustainability [4]. In this paper we discuss our research focussed on application of statistical methods, particularly non-classical approaches, for describing dynamics within land as a coupled human-environment system. We use long time-series of historical data to test and develop methods and methodology, as well as interpret dynamics in land systems over different timespans. Our perspective is that, as a coupled human-environment system, land responds to pressures and drivers from social, economic, political, technological, and environmental factors, and draws on social, human, economic, technical, and natural capitals, to produce flows of goods (and "bads") from these capital funds. Within this attention to whole system issues and whole system dynamics, we argue that sustainability is most usefully understood as a property of both the dynamic and dynamical nature of the coupled system. As such, its measurement must address sustainability as an emergent property of the functional dynamics of the whole coupled human-environment system over different timescales simultaneously.

2 Materials and Methods

Data are from the annual agricultural census from 1867-2020 for Scotland. These are a national aggregate summary for the entire period, and for 33 counties from 1876-1975, the county administrative structure being replaced from 1976 onwards. Data include arable area and livestock counts, including for cereals and sheep. Previous work has investigated multi-time scale dynamics for the national aggregate data from 1867-2020, identifying both endogenous dynamics in cereals and sheep as emergent properties of the farming system as it has responded to international and national market prices for cereals and also impacts of exogenous factors such as weather and disease [2]. The national aggregate data also have been used to track dynamic changes in provisioning ecosystem services and modernisation of agriculture from 1940-2016 [1]. These papers contain full details of the data used [1, 2]. Analysis uses time series analysis.

3 Results

Figure 1 shows some measures of inputs for agriculture, and indices of resource use efficiency and economic return for farming in Scotland from 1940 to 2016. These figures are modified from the flow-fund accounting methods identified in [1]. The figures show that although farming has become more efficient in conversion of financial resources into food over time (food production conversion efficiency), it has become less efficient in use of natural resources, requiring increasing inputs of energy (HP) and Nitrogen fertiliser applied, and the economic return from food production has declined (also leading to declining social capital in farming). These provide a picture of post-war dynamics of farming in Scotland that is in no way sustainable, in economic, social or environmental terms. This is neither new nor surprising, but the time series provide a laboratory for investigation of dynamics that offers scope for measurement of sustainability (and "unsustainability").

Figure 2 shows the long-term trends and medium-term cycles for cereal area and sheep numbers over a 100-year period for 33 county areas in Scotland, as well as for Scotland as a whole. Long-term trends follow different paths of increase and decrease among the counties, yet all counties exhibit similar medium-term dynamics, namely irregular cycles with about 30-year period. Previous research has shown these medium-term cycles to be linked to cycles in national and global prices of cereals and their coupled dynamics to vary in relation to different periods of stability in government agricultural policies [2]; the cycles are interpreted as embedded damping of variability in farming systems across Scotland associated with endogenous system dynamics.

4 Discussion

The results provide not only a range of meaningful flow, fund, and flow-fund indicators that can be derived from agricultural data, but also long-term and medium-

term trends that can be interpreted using time series analysis. Further analysis of the dynamics represented by these data and indicators can be developed using Recurrence Plots [10] and Recurrence Quantification Analysis [13] used for study of non-linear dynamical systems. Indicators have been found to be useful for assessment of resilience of agroecosystems [5, 12], and fund, flow and fund-flow ratios have been used in a multiscale accounting methodology for sustainability assessment [9], combining funds and flows of water, energy, food, and human activity, with land use and economics, at multiple spatial and temporal scales to evaluate the feasibility, viability, and desirability of demand and supply for the connected funds and flows. Our approach is consistent with the logic of this "whole system" resource accounting methodology.

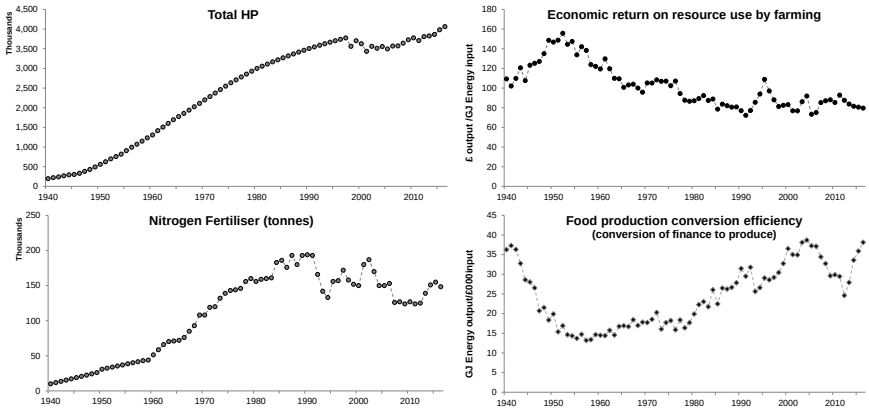


Figure 1. Indicators of farming efficiency for Scotland from 1940-2016: a) Use of energy (total HP available from human, animal, and mechanised tractor work), b) tonnes of nitrogen fertiliser applied, c) economic return on resource use by farming (£output/GJ energy input), d) food production conversion efficiency (GJ energy output/£000 input).

Further measurement and analysis of trends, structures and interactions of funds, flows, and fund-flow ratios over time and space, as well as influences of endogenous system forces and exogenous factors, offers scope for adding understanding of system dynamics over time, developing the application of non-standard methods for time series analysis to address dynamical systems and their resilience and sustainability.

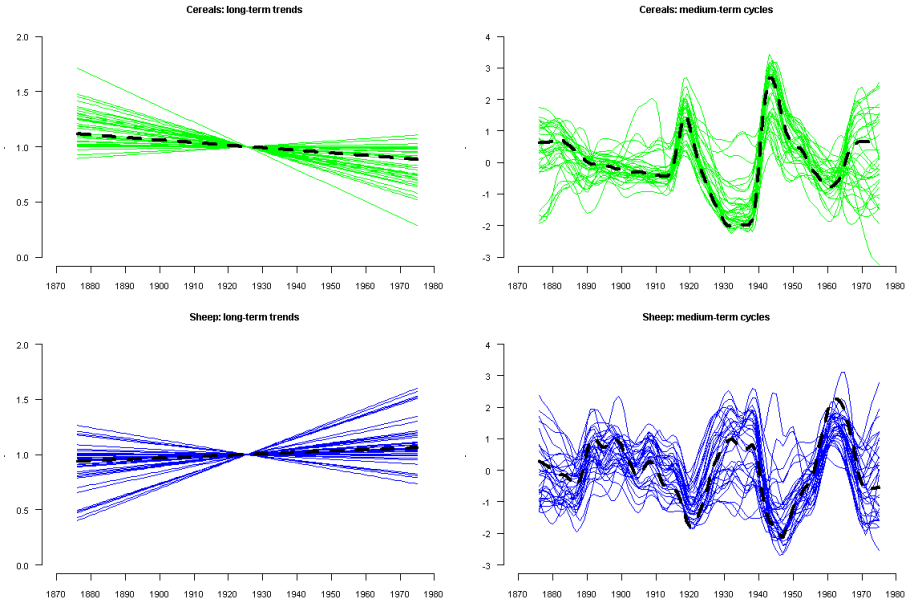


Figure 2. Long-term trends and medium-term cycles for cereal area and sheep numbers in 33 counties of Scotland, 1876-1975. The dashed lines are the whole Scotland trends and cycles for each variable.

References

1. Aspinall, R.J. Staiano, M.: Ecosystem services as the products of land system dynamics: lessons from a longitudinal study of coupled human–environment systems. Landscape Ecology **34**(7), 1503-1524 (2019)
2. Aspinall, R.J., Staiano, M.Pearson, D.M.: Emergent Properties of Land Systems: nonlinear dynamics of Scottish farming systems from 1867 to 2020. LAND (2021)
3. Benton, T.G., Bailey, R., Froggatt, A., King, R., Lee, B.Wellesley, L.: Designing sustainable landuse in a 1.5°C world: the complexities of projecting multiple ecosystem services from land. Current Opinion in Environmental Sustainability **31**, 88-95 (2018)
4. Ehrensperger, A., de Bremond, A., Providoli, I.Messerli, P.: Land system science and the 2030 agenda: exploring knowledge that supports sustainability transformation. Current Opinion in Environmental Sustainability **38**, 68-76 (2019)
5. Eichler Inwood, S.E., López-Ridaura, S., Kline, K.L., Gérard, B., Monsalue, A.G., Govaerts, B.Dale, V.H.: Assessing sustainability in agricultural landscapes: a review of approaches. Environmental Reviews **26**(3), 299-315 (2018)
6. FAO: Climate-smart agriculture case studies 2021 - Projects from around the world.: Rome. (2021)
7. Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty,

N.Snyder, P.K.: Global consequences of land use. Science **309**(5734), 570-574 (2005)

8. Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C., Balzer, C., Bennett, E.M., Carpenter, S.R., Hill, J., Monfreda, C., Polasky, S., Rockstrom, J., Sheehan, J., Siebert, S., Tilman, D.Zaks, D.P.M.: Solutions for a cultivated planet. Nature **478**(7369), 337-342 (2011)
9. Giampietro, M., Aspinall, R., Ramos-Martin, J.Bukkens, S.: Resource Accounting for Sustainability Assessment: the nexus between energy, food, water and land use. Explorations in Sustainability and Governance Series. Routledge (2014)
10. Marwan, N., Carmen Romano, M., Thiel, M.Kurths, J.: Recurrence plots for the analysis of complex systems. Physics Reports **438**(5), 237-329 (2007)
11. Ritchie, H. Roser, M.: Land Use., Published online at OurWorldInData.org. Retrieved from www.ourworldindata.org/land-use [Online Resource]. (2019)
12. van Apeldoorn, D.F., Kok, K., Sonneveld, M.P.W.Veldkamp, T.: Panarchy Rules: Rethinking Resilience of Agroecosystems, Evidence from Dutch Dairy-Farming. Ecology and Society **16**(1) (2011)
13. Webber, C.L. Marwan, N., eds. Recurrence Quantification Analysis: Theory and Best Practices. Understanding Complex Systems. Springer: London. 436 (2015)

Tourism sustainability in the Italian regions: a fuzzy clustering approach

Sostenibilità turistica nelle regioni italiane: una classificazione fuzzy

Leonardo Salvatore Alaimo and Giovanni Finocchiaro

Abstract Impact of tourism on the environment is rarely object of investigations and in-depth studies because of the lack of appropriate official statistics. Taking into account the annual time series of the ISPRA environmental indicators (2015-2019) selected to measure the sustainability of tourism, we try to investigate this phenomenon in the Italian regions. Using a fuzzy clustering approach, two clusters of Italian regions were identified. Regions of the two groups are characterised for the different environmental pressure suffered by their territories and the use of their accommodation offer.

Abstract *Tenendo conto delle serie storiche annuali (2015-2019) degli indicatori ambientali selezionati da ISPRA per misurare la sostenibilità del turismo, questo lavoro cerca di indagare la sostenibilità turistica delle regioni italiane. L'applicazione di una fuzzy cluster analysis ha portato all'identificazione di 2 cluster omogenei di regioni italiane, caratterizzate da una maggiore o minore pressione ambientale subita dai territori e da un migliore o peggiore utilizzo della propria offerta ricettiva.*

Key words: Tourism, Sustainability, Italian regions, Fuzzy clustering, Time series

1 Introduction

Tourism and the environment represent a fundamental combination for many world tourist destinations as, often, it is the natural heritage that makes them attractive.

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

Giovanni Finocchiaro
ISPRA – Italian Institute for Environment Protection and Research, Via Vitaliano Brancati, 48, 00144 Rome, Italy, e-mail: giovanni.finocchiaro@isprambiente.it

However, it is not as clear how the environment should not be damaged by tourism or for tourism. The monitoring of the relationship between tourism and the environment is necessary, in particular, for the control of the impacts of tourism on the environment and, consequently, for the implementation of policies aimed at ensuring greater tourism sustainability in the various territories. The relationship between tourism and the environment is essentially "unknown" to official national and European statistics. Tourism is generally seen as a productive sector dedicated to creating income, and the statistics available on tourism are essentially designed to measure the economic role of tourism, whereas its effects on the environment are not really systematically measured [6, 3]. However, the environmental impact of tourism is receiving more and more attention in the general framework of Sustainable Development Goals (SDGs) and Agenda 2030. At a global level, tourism, although it has been explicitly linked to the 8, 12 and 14 Goals of the 2030 Agenda (SDGs), has the potential to contribute, directly or indirectly, to all of them. Indeed, the United Nations World Tourism Organization (UNWTO) has defined how each objective corresponds to a specific response from the tourism sector. This work devotes its attention to three mentioned SDGs by proposing a measure to monitor the environmental effects of tourism in a integrative way, i.e. trying to give an integrated reading of the environmental indicators produced in Italy by ISPRA (the only institutional reporting experience on tourism and the environment). The work therefore seeks to provide an integrated reading of environmental indicators to measure the current sustainability of the tourism sector, much desired from an environmental point of view and just as little considered from an economic point of view. The focus is on the Italian regions, given the profound territorial differences that characterise the territory. The aim of this paper is to identify homogeneous clusters of Italian regions, taking into account the yearly time series of the indicators selected to measure the sustainability of tourism, in the period 2015-2019. In this way, we analyse the evolution over time of this concept in the Italian regions, finding hidden patterns or similar groups and highlighting their characteristics. So, we can examine the evolution over time of tourism sustainability and take into account its territorial characteristics. In order to deal with the complexity and uncertainty of this concept, we adopt a fuzzy approach. 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.

2 Data description and methods

Table 1 reports the indicators selected for this work and their description. We must underline that all variables present negative polarity with respect to tourism sustainability (i.e., the higher their values, the worst the situation in terms of tourism sustainability), except "X5 - Tourism infrastructures" that has positive polarity. All the indicators have been normalised using a Min-Max normalisation.

Table 1: Indicators of tourism sustainability; code; variable name; description.

Code	Variable name	Description
X1	Tourism intensity	Ratio between the overnight stays (number of nights spent in the area by those who practice tourism) and the population leaving in the area.
X2	Nox Emissions produced by road travel of Italian residents on Italian territory for tourism purposes	Tons of nitrogen oxides (Nox) produced by road travel by Italian residents for tourism purposes.
X3	Electricity consumption in tourism sector	Electricity consumption for the NACE Groups 55.3, 55.4 and 55.5: "Hotels, restaurants and bars".
X4	Quote of municipal waste per capita attributable to tourism	Per capita share of urban waste attributable to tourism.
X5	Tourism infrastructures	Ratio between the number of nights spent in the area by those who practice tourism, recorded in the hotels, and the product of the number of opening days of the hotels by the number of beds.

In this paper, we use a fuzzy clustering method for multivariate time series based on the so-called Dynamic Time Warping (DTW) distance [9, 2]. It is based on the dilatation or contraction of two (multivariate) time series locally, in order to make their shape as similar as possible. The total distance between two time series $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ is computed through the so called "warping path", which ensures that each data point in $\mathbf{X}_i$ is compared to the "closest" data point in $\mathbf{X}_{i'}$. Let:

$$\Phi_l = (\varphi_l, \psi_l),\ l = 1, \ldots, L. \tag{1}$$

under the following constraints:

1. boundary condition: $\Phi_1 = (1,1),\ \Phi_L = (T, T')$;
2. monotonicity condition: $\varphi_1 \leq \ldots \leq \varphi_l \leq \ldots \leq \varphi_L$ and $\psi_1 \leq \ldots \leq \psi_l \leq \ldots \leq \psi_L$.

The warping curve realigns the time indices of $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ through the functions φ and ψ. The total dissimilarity between the two "warped" multivariate time series is:

$$\sum_{l=1}^{L} d(\mathbf{x}_{i,\varphi_l}, \mathbf{x}_{i',\psi_l}) m_{l,\Phi} \tag{2}$$

where $m_{l,\Phi}$ is a local weighting coefficient; $d(.,.)$ is, usually, the Euclidean distance for multivariate time series:

$$d(i, i') = (\|\mathbf{x}_{it} - \mathbf{x}_{i't'}\|)^{\frac{1}{2}}. \tag{3}$$

The DTW distance is the one which corresponds to the optimal warping curve among the several warping curves, $\hat{\Phi}_l = (\hat{\varphi}_l, \hat{\psi}_l),\ l = 1, \ldots, L$ which minimizes the total dissimilarity between $\mathbf{X}_i$ and $\mathbf{X}_{i'}$:

$$d_{DTW}(\mathbf{X}_i, \mathbf{X}_{i'}) = \sum_{l=1}^{L} d(\mathbf{x}_{i,\hat{\varphi}_l}, \mathbf{x}_{i',\hat{\psi}_l}) m_{l,\hat{\Phi}}. \tag{4}$$

The following exponential transformation of the DTW distance is used:

$$_{exp}d^2_{DTW}(\mathbf{X}_i,\mathbf{X}_{i'}) = 1-\exp\left\{-\beta\, d^2_{DTW}(\mathbf{X}_i,\mathbf{X}_i)\right\} \tag{5}$$

where β is a suitable parameter (positive constant) determined according to the variability of the data [5]. In our study, we use the Dynamic Time Warping-based Fuzzy *C*-Medoids clustering model with Exponential transformation (DTW-Exp-FCMd) (for details, please see: [5]):

$$\begin{cases} \min: & \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \; {}_{exp}d^2_{DTW}(\mathbf{X}_i,\widetilde{\mathbf{X}}_c) = \\ & \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \left[1-\exp\left\{-\beta\, d^2_{DTW}(\mathbf{X}_i,\widetilde{\mathbf{X}}_c)\right\}\right] \\ s.t. & \sum_{c=1}^{C} u_{ic} = 1,\; u_{ic} \geq 0 \end{cases} \tag{6}$$

For choosing the optimal partition, *C*, we use the Xie-Beni criterion [8]:

$$\min_{C\in\Omega_C}: I_{XB} = \frac{\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ik}^m d^2_{DTW}(\mathbf{X}_i,\widetilde{\mathbf{X}}_c)}{I\min_{c,c'} d^2_{DTW}(\widetilde{\mathbf{X}}_c,\widetilde{\mathbf{X}}_{c'})} \tag{7}$$

where Ω_C represents the set of possible values of $C\,(C<I)$. The optimal number of clusters *C* is identified in correspondence with the lower value of I_XB (for more detail, please see: [4]). The used time series clustering approach presents all the general advantages connected to the fuzzy theory in a clustering framework, is sensitive in capturing the dynamic characteristics of the time series and inherits the advantage connected to DTW distance (for details, please see: [4]). Furthermore, being an observation-based clustering method, it is particularly suitable for classifying short time series, as those treated in this paper (for details, please see: [1, 4]).

3 Application and results

We compute the Xie-Beni index for $2 \leq c \leq 4$ obtaining these results: $I_{XB}(c=2) = 0.506$; $I_{XB}(c=3) = 0.796$; $I_{XB}(c=4) = 1.332$. According to the criterion, we chose the 2 clusters partition. For the evaluation of the fuzziness, we need to specify a cut-off point for the membership degree. If we have a two-clusters situation and the membership degrees in both clusters are between 0.3 and 0.7, it would be considered that there is a reasonable level of fuzziness in the cluster membership of the time series (for more information on the choice of cut-off, please see: [4, 7]). The solution identifies two medoids, Emilia Romagna (EMR) for cluster 1 and Marche (MAR) for cluster 2, and one fuzzy region, Apulia (PUG).

Figure 1 shows the subdivision of the Italian regions according to the cluster to which they belong and the matrix with the membership degrees. The regions' membership to their respective clusters, apart from the fuzzy region, is clear and unambiguous. There is no split of the country between northern and southern regions. Clusters are clearly characterised; the first (including 7 regions) presents

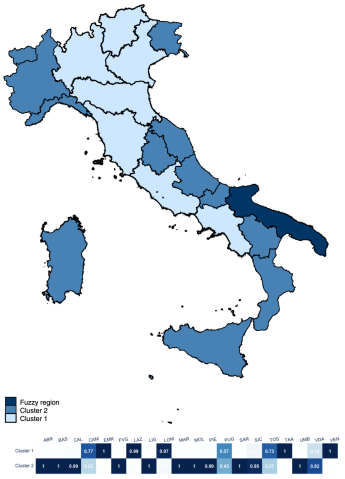


Fig. 1 Tourism sustainability: clusters' composition and membership degrees.

higher trends than Italy, while the second (including 12 regions) lower ones. This is clearly shown in Figure 2, which presents the comparison between the two clusters' medoids, respectively EMR and MAR, Italy and the fuzzy region, Apulia (PUG). Cluster 1 includes the regions that have more negative externalises than the national data from the environmental point of view, but at the same time they have the tourist infrastructures better exploited and probably better managed. Cluster 2 presents the exact opposite situation. PUG clearly presents different trends than those of the two medoids. If, in fact, it presents trends better than Cluster 2 in tourism intensity and quote of municipal waste per capita (indicating a better situation in terms of tourism sustainability), at the same time it has trends worse than Cluster 1 in tourism infrastructures and Nox emissions. Regarding the electricity consumption, it has a trend very similar to the national one.

4 Conclusions

Given the current absence at European level of official statistical surveys aimed at collecting data useful for monitoring the relationship between tourism and the environment, we try to approach this analysis by using the environmental indicators of ISPRA about "Tourism & Environment". Our analysis, conducted for the Italian regions yearly time series (2015-2019), aims at identifying homogeneous clusters. For this purpose, we adopted a clustering method particularly suitable for the data analysed. We identified two specific clusters of Italian regions, not grouped by geographic origin (North-South), but by their way of managing and/or experiencing tourism in their territories. In detail, the regions belonging to cluster 1, that is the ones that show values exceeding the national ones, are 6 out of 8 among the most populous in Italy and also host the largest Italian metropolitan cities. So cities al-

Fig. 2: Tourism sustainability: comparison among Cluster 1, Cluster 2, Italy and Apulia.

ready in themselves with an imposing endogenous and structural demographic pressure that inevitably increases with tourism. Therefore, those regions should be more careful than others to ensure their territories a more sustainable approach than the tourism they offer.

References

1. Alaimo, L.S.: Complexity of social phenomena: Measurements, analysis, representations and synthesis. Unpublished doctoral dissertation, University of Rome "La Sapienza", Rome, Italy, (2020).
2. Berndt, D.J., Clifford, J.: Robust Fuzzy Clustering of Multivariate Time Trajectories. Proceedings of the AAAI-94 Workshop Knowledge Discovery in Databases. 359–370 (1994).
3. Betta, L., Dattilo, B., di Bella, E., Finocchiaro, G., Iaccarino, S.: Tourism and Road Transport Emissions in Italy. Sustainability **3**, 12712 (2021). doi: 10.3390/su132212712.
4. D'Urso, P., Alaimo, L.S., De Giovanni, L., Massari, R.: Well-Being in the Italian Regions Over Time. Soc. Ind. Res. (2020). doi: 10.1007/s11205-020-02384-x.
5. D'Urso, P., De Giovanni, L., Massari, R.: Using Dynamic Time Warping to Find Patterns in Time Series. Int. Jou. App. Rea. **99**, 12–38 (2018).
6. Finocchiaro, G., Iaccarino, S., Salomone, M.: Ambiente: Sfida e opportunità per il turismo. ISPRA – Stato dell'ambiente 73/2017 – ISBN:978-88-448-0826-6 . (2017).
7. Maharaj, E.A., D'Urso, P.: Fuzzy Clustering of Time Series in the Frequency Domain. Inf. Sci. **181(7)**, 1187–1211 (2011).
8. Xie, X.L., Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis & Machine Intelligence. **13(8)**, 841–847 (1991).
9. Velichko, V.M., Zagoruyko, N.G.: Automatic Recognition of 200 Words. Int. Jou. Man. Stu. **2(3)**, 223–234 (1970).

How young people perceive environmental issues, react to ecological concerns and commit themselves to sustainable behaviours

Come i giovani percepiscono le questioni ambientali, reagiscono alle preoccupazioni e si impegnano per adottare comportamenti sostenibili

Angela Maria D'Uggento

Abstract Environmental issues have gained widespread prominence, contributing to growing environmental awareness and positive action for the environment. Young people have participated in global green movements that call on policy makers to consider environmental protection as a priority on their agendas. This paper aims to investigate young people's awareness of the dangerous effects of climate change and subsequent pro-environmental actions through an online survey of *2,041* high school students. Data analysis is conducted using Classification Trees and Random Forest to identify sustainable behaviours. The results show that the younger generation has a strong awareness of environmental conditions and is, in part, engaged by adopting best ecological practices and actively caring for the environment.

Abstract *Le questioni ambientali hanno acquisito un'importanza sempre maggiore, contribuendo alla crescita di una coscienza ambientale e di azioni positive per l'ambiente. I giovani partecipano a movimenti verdi di rilevanza planetaria sollecitando i governanti ad inserire la protezione ambientale tra le priorità nelle loro agende. Il contributo intende valutare la consapevolezza dei giovani sugli effetti pericolosi del cambiamento climatico e le conseguenti azioni a favore dell'ambiente attraverso un sondaggio online su 2,041 studenti delle scuole superiori. È stata effettuata una classificazione ad albero ed una Random Forest per identificare i principali comportamenti virtuosi sostenibili per l'ambiente. I risultati mostrano che le giovani generazioni hanno una forte consapevolezza delle condizioni ambientali e in parte, si impegnano adottando best practices ecologiche.*

[1] Angela Maria D'Uggento, Università degli Studi di Bari Aldo Moro; angelamaria.duggento@uniba.it

Key words: environmental problems, sustainable behaviours, Classification and Regression Trees, Random Forest.

1 Introduction

Climate change is the most important global crisis of our time, causing increasingly serious damage to our planet. It is a complex but solvable challenge that can be an opportunity for a prosperous and resilient economy. Institutions, businesses and citizens are called upon to contribute, according to their capabilities and roles. People today are aware of the urgent need to take positive action to protect the environment and promote sustainable economic development by adopting virtuous behaviours that form the pillars. Since adults put the world in the hands of young people, this awareness should begin at a young age and continue to be one of the basic principles of personality in adulthood. In the early stages of life, the main educators, i.e. family and school, can play an important role in environmental education. The aim of this paper is to investigate whether young people simply share the ethical statements on environmental protection, driven by social desirability, or whether they are protagonists of sustainable behaviours that are part of their lifestyle. The survey was conducted in 2018. At that time, Fridays for Future (FFF) established itself as an international protest environmental movement composed of students who participated in demonstrations around the world demanding action to prevent global warming and climate change. The goal of the FFF strikes is to draw attention to the issue of climate change in order to make it a priority on the international political agenda. Against this backdrop, all those who wish for a more sustainable future are pinning their hopes on youth to reverse the damaging course of recent years. But, do young people really have a strong environmental awareness and do they engage in sustainability on a daily basis and with concrete actions? Answering these questions could help identify sustainable behaviours suitable for spreading best practices among young people through the most effective means, i.e., educational institutions or volunteer associations.

The data collected in a survey (in the framework of the Italian Ministry of Education's National Project for a Scientific Degree in Statistics - PLS) involving 2,041 students of Apulian high schools (southern Italy) provided interesting findings. In particular, the paper aims to understand the extent to which young people care about the environment and whether they really feel the seriousness of the situation and intend to commit themselves to a more sustainable world. From the findings, useful policy guidelines can be derived for disseminating best practices that the entire community should follow to adopt sustainable behaviours. The paper is organized as follows: after the introduction, section two deals with a description of the survey, data and methods; section three with the main results of the statistical analysis and discussion; section four concludes with some brief remarks.

2 The survey, data and methods

Student participation in the survey was on a voluntary basis, with anonymity assured by formal privacy consent. They were given a questionnaire with 28 questions divided into three sections: I. Knowledge of the phenomenon and main concerns about environmental problems; II. Sustainable behaviours and lifestyle; III. Future expectations and suggestions on the role of institutions and citizens. Students were asked to express their opinions and perceptions about environmental issues using a five-point Likert scale, with 1 being the lowest, 5 being the highest, and 3 being considered a neutral response. After the exploratory analysis, which was useful to understand the phenomenon from the adolescents' point of view, some variables were selected for a more in-depth analysis through Classification trees (Breiman *et al.*, 1984) and the Random Forest (Breiman, 2001). Classification trees allow hierarchical segmentation of a large group of individuals and identification of patterns based on some selected variables. They are widely used in data mining studies because they are more flexible and do not require strong assumptions about the distribution of the dependent variable which can also be categorical, as is often the case in social studies. In general, this technique is one of the most popular supervised machine learning algorithms used for classifications. Then, the Random Forest (RF) method is used to evaluate and rank variables in terms of their ability to predict the response by means of the variable importance measures (VIMs). Given an error measure M (e.g., error rate or mean squared error), VIM is defined as:

$$VIM_j^M = \frac{1}{ntree}\sum_{t=1}^{ntree}\left(MP_{tj} - M_{tj}\right)$$

where *ntree* is the number of trees in the forest; MPtj indicates the error of the tree t when predicting all observations that are out-of-bag for tree t after randomly permuting the values of the j-th predictor variable. Similarly, Mtj indicates the above-mentioned error of the tree t before permuting the values of the j-th predictor variable. The RF method has some important advantages: it is not parametric, it is not based on a particular stochastic model since no specific distribution of the response variable is assumed, and it does not require specification of the type of relationship (linear or nonlinear) between the response variable and the predictors. Finally, it allows for more accurate results in terms of a more robust assessment of the importance of the variable compared to classical tree-based methods. In the corresponding analyses, the dependent variable "Contribution to environmental protection" is dichotomous and takes the value "None" if the respondent declared that he/she is not willing to contribute to environmental protection or "Personally" if he/she wants to play an active role. The selected predictors used in Classification tree and RF are: Gender, Oil recycling, Battery recycling, Drugs recycling, Electric cable recycling, Volunteer, Paper sheets reuse, Water tap switch off, Standby Off,

Shopping bags reuse. Classification tree were performed using the software IBM SPSS and Random Forest using the packages *randomForest* in environment R.

3. Main results and discussion

3.1 Some insights from exploratory analysis

The aim of the analysis is to detect students' attitudes towards environmental issues, perceptions of the seriousness of the main problems such as global warming, air and water pollution, depletion of natural resources, deforestation and then the contribution to the protection of the environment through their daily behaviour and their vision of the future. The main statistics summarizing the students' responses and their characteristics are shown in Table 1.

Table 1: Statistics on the main items under investigation

Variables						
Gender (% Female)	53.2					
Age (mean/std)	16.3	1.5				
Environmental problems/ratings	1	2	3	4	5	mean of ratings
Water pollution(%)	7.3	13.5	22.3	24.1	32.8	3.6
Air pollution(%)	2.4	9.0	17.8	28.6	42.3	4.0
Depletion of natural resources (%)	4.5	11.5	22.8	27.1	34.1	3.7
Global warming (%)	3.4	9.8	18.1	29.2	39.4	3.9
Deforestation (%)	5.0	9.0	19.6	30.0	36.4	3.8
Noise pollution (%)	7.9	19.8	32.1	26.6	13.6	3.2
Recycling (%Yes)	83.0	Oil recycling (%Yes)				40.9
Plastic recycling (%Yes)	98.8	Battery recycling (%Yes)				55.7
Glass recycling (%Yes)	94.0	Drugs recycling (%Yes)				57.8
Organic recycling (%Yes)	87.4	Electric cable recycling (%Yes)				31.0

3.2 Classification tree and Random Forest to detect young people's sustainable behaviours

The data show that students are very concerned about the future of the planet, but confident that we can still intervene because the problems are not unsolvable. They report being conscientious about all types of recycling required by law (glass, plastic, paper, organic waste), but this should be taken for granted and the corresponding variable may not be an effective predictor of environmental sensitivity. On the other hand, if we focus our attention on the so-called "non-mandatory" behaviours, that depend solely on the will of the individual, we can extract the significant variables that make the difference by having a positive impact on environmental protection.

Figure 1 below shows the Classification tree. The first model developed included all of the selected predictors but was not considered to be very informative. Consequently, only the variables related to voluntary behaviour and gender were entered. The response variable focuses on the willingness to play an active role in the defence of the planet as an individual; the students who declared to engage and spread virtuous behaviours represent only 42.7% of the respondents. The first explanatory variable that allows us to profile the "very committed environmentalists" is the recycling of batteries. This undoubtedly proves to be a good predictor of the presence of ecological awareness, since batteries are among the most polluting materials and must be disposed of in special containers that are not distributed in the area as frequently as those for glass, paper and plastic. In addition, these students (55.2% of respondents) are more willing to volunteer (cleaning beaches, public places, local entertainment) and show greater awareness against the waste of precious natural resources, both in simple daily activities such as turning off TV/PC/devices or the faucet and in more heavy ones, such as the disposal of oil and power cables, both of which must be transported to the eco-island. The alternative path in the tree after the “Battery recycling split” is probably more interesting as it allows us to identify those who, with the right incentives and not so much effort, could become "very committed environmentalists". We might, instead, call them "most comfortable environmentalists" because these students still engage in basic sustainable behaviours, such as almost always reusing grocery bags and sheets of paper, but do not pursue more demanding activities. In a few cases, they may participate in volunteer activities and not waste water. They tend to believe that it is enough to follow the guidelines prescribed by law, such as separating waste by glass, paper and plastic, thus conforming to socially desirable statements.

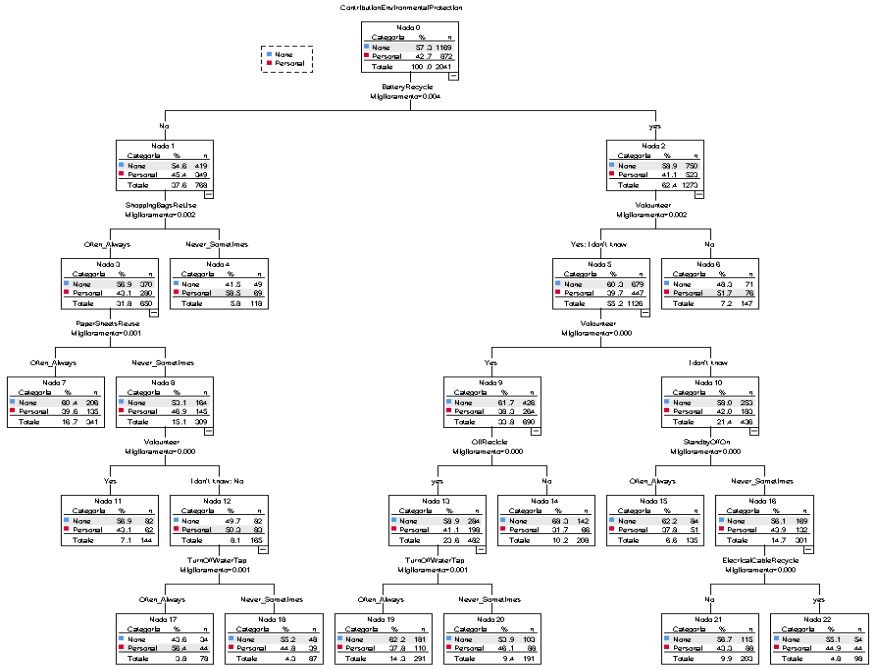
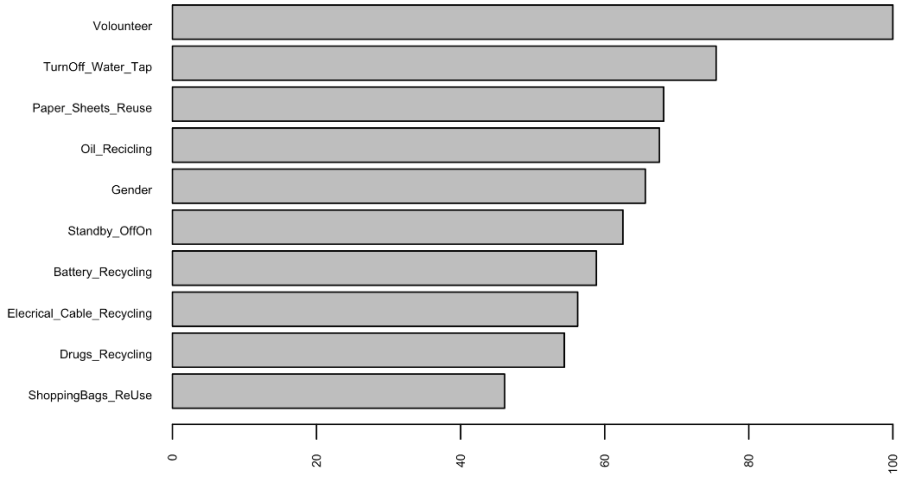


Figure 1: Classification tree for Contribution to environmental protection. Overall percentage of correct classification: 59.0 (None 83.2; Personal 16.8) Risk estimate: 0.410.

A RF with 5000 trees was created. Fig.2 shows that the most important variables are closely related to behaviours that strongly indicate a commitment to the environment, since they are not mandatory but entrusted to the will of the individual and therefore represent good practices in daily life. These behaviours indicate the levers to pull in order for them to become good practices in everyday life.



How young people perceive environmental issues, react to ecological concerns and commit themselves to sustainable behaviours
Figure 2: Importance of predictors in Random Forest

4 Some brief remarks

Young people are aware of the seriousness of environmental problems and are confident that they can still be solved if they are addressed urgently. However, some of them need to be encouraged to get personally involved so that sustainable behaviours can become part of the culture of our communities. People tend to conform to social desirability, but their environmental awareness needs to translate into best practices for sustainability. The most effective way should be found to get the "comfortable" environmentalists to become more engaged, even if it is not easy to make a direct connection between values and action. A fundamental role in this direction can be played by actors outside the family, such as volunteer associations, which are composed of young people and can suggest sustainable models that can be emulated by their peers. It is critical to capture their interest by using compelling messages and appropriate language and tools to sustain it. Schools can have a positive impact on young people's environmental education, but 83.0 of respondents said that teachers pay little attention to the environment. Therefore, educational institutions need to give more consideration to the debate on environmental issues. It is likely that among the critical issues highlighted by the Covid19 pandemic, environmental protection has continued to gain priority.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Belmont, CA (1984)
2. Breiman, L.: Random forests. Machine Learning **45(1)**, 5–32 (2001)

Session of solicited contributes SS16 – *Statistical Methods for Environmental, Natural Resources and Health Assessment*
Organizer and Chair: Alessio Pollice

Spatial-ARFIMA models for the statistical analysis of environmental lattice processes

Modelli ARFIMA per l'analisi statistica di processi ambientali osservati su lattice

Angela Ferretti, Luigi Ippoliti and Pasquale, Valentini

Abstract Long memory is playing an important role in many scientific disciplines and applied fields such as physics, hydrology, economics, biology, telecommunications and environmental sciences. In this work we describe a class of anisotropic stationary lattice processes with long memory which represents an extension in two-dimensions of the ARFIMA models widely used in time series. Preliminary analysis of satellite data on See Surface Temperature are also considered.

Abstract *Abstract in Italian* Diversi modelli basati sul concetto di lunga memoria occupano un ruolo importante in molte discipline scientifiche quali la fisica, l'idrologia, l'economia, la biologia, le telecomunicazioni e le scienze ambientali. In questo lavoro descriviamo una classe di processi spaziali anisotropici con lunga memoria che rappresenta un'estensione in due dimensioni dei modelli ARFIMA ampiamente utilizzati nelle serie temporali. Sono inoltre considerate alcune analisi preliminari di dati satellitari riguardanti la temperatura della superficie del mare.

Key words: Long-range dependence, Spatial ARFIMA models, See surface temperature, Environmental processes

Angela Ferretti
University G.d'Annunzio, Viale Pindaro 42, 65127 Pescara, Italy, e-mail: angelaferretti29@gmail.it

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

Pasquale Valentini
University G.d'Annunzio, Viale Pindaro 42, 65127 Pescara, Italy, e-mail: pasquale.valentini@unich.it

1 Long memory random fields

In many areas of applications, such as ecology, environmental monitoring and remote sensing, data are observed on a regular lattice. A standard assumption for many random fields proposed to model such data usually assumes that the spatial correlation decays exponentially with the distance. In some cases, however, it has been observed that the spatial dependence does not decay quickly enough, and the correlation, though small, does not appear negligible even for distant observations. This was found in many scientific disciplines and applied fields such as environmental sciences [1, 2], image analysis [3] and agricultural fields [4, 5, 6].

Because of the slow decay of the correlation function, these processes are usually referred to as *long memory* (LM) or *long range dependent* (LRD) processes. Long-range dependence may be defined in many ways. However, because second order properties of a stochastic process are well known concepts, definitions based on the behaviour of covariances and the spectral density are by far the most popular found in literature.

Let $Y = \{Y_{s,t},\ (s,t) \in \mathbb{Z}^2\}$ be a weakly stationary real-valued random field, with mean zero, autocovariance function $R_y(u,v) = \mathrm{Cov}\,(Y_{s,t}, Y_{s+u,t+v})$, and autocorrelation function $r_y(u,v) = R_y(u,v)/\sigma_y^2$, where $\sigma_y^2 = R_y(0,0)$. Furthermore, for a pair of spatial frequencies $(\lambda_1,\lambda_2) \in (-\pi,\pi]^2$, let

$$f_y(\lambda_1,\lambda_2) = (2\pi)^{-2} \sum_{(u,v)\in\mathbb{Z}^2} R_y(u,v) e^{-iu\lambda_1} e^{-iv\lambda_2}$$

be the spectral density function of Y. We say that the random field is a long memory stationary processes if its autocovariance function decays hyperbolically. In particular, decaying as a power law in the lag variable, the covariances sum to infinity for large lag values, that is

$$\sum_{(u,v)\in\mathbb{Z}^2} |R_y(u,v)| = \infty.$$

Alternatively, a random field with unbounded spectral density at certain frequencies (including zero frequency) is considered to exhibit long-range dependence. These time- and frequency-domain characterizations of long-range dependence are closely related, but not equivalent [7]. In this work, the frequency-domain characterization is employed for our statistical analysis. Although analysing the variability of a process via the covariance function and the spectral density can be regarded as equivalent, they provide different ways of analysing the process, and spectral analysis offers some computational advantages compared to analysis based on the covariance function.

The remainder of this short paper is organized as follows. In section 2, we introduce the spatial autoregressive fractional integrated moving average (ARFIMA) model while in section 3 we provide a preliminary exploratory analysis of dataset on See Surface Temperature over a region of the Pacific Ocean.

2 Spatial ARFIMA models

In this section we introduce the Spatial ARFIMA model for a class of anisotropic stationary lattice processes with long memory. Based on the use of two long-memory parameters, d_1 and d_2, in the horizontal and in the vertical directions, respectively, the model represents an extension of the separable ARIMA process introduced by [8] for short memory spatial processes. The model fits nicely into the Box-Jenkins modelling framework and it is thus computationally appealing - see, for example, [9, 10].

Assume that Y is a Gaussian process defined on a finite two dimensional regular rectangular lattice $\mathscr{L}$ with $n = n_1 \times n_2$ sites. Also, let L_1 and L_2 be backward shift operators on $\mathbb{Z}^2$ with $L_1\, Y_{s,t} = Y_{s-1,t}$ and $L_2\, Y_{s,t} = Y_{s,t-1}, (s,t) \in \mathbb{Z}^2$. Finally, assume that $X = \{X_{s,t},\ (s,t) \in \mathbb{Z}^2\}$ is a stationary, invertible and separable ARMA process [8] satisfying

$$A(L_1,L_2)\, X_{s,t} = B(L_1,L_2)\, \eta_{s,t}$$

where $\{\eta_{s,t}\}$ is a zero mean white noise process with variance $\sigma_\eta^2 < \infty$. Because of separability assumption it also follows that

$$A(L_1,L_2) = A_1(L_1)\, A_2(L_2) \qquad\qquad B(L_1,L_2) = B_1(L_1)\, B_2(L_2)$$

where

$$A_1(L_1) = 1 - \sum_{u=1}^{p_1} \alpha_{1u} z^u, \qquad\qquad A_2(L_2) = 1 - \sum_{v=1}^{p_2} \alpha_{2v} z^v$$

$$B_1(L_1) = 1 + \sum_{u=1}^{q_1} \beta_{1u} z^u, \qquad\qquad B_2(L_2) = 1 + \sum_{v=1}^{q_2} \beta_{2v} z^v.$$

with (p_1,q_1) and (p_2,q_2) being the order of the ARMA model in the horizontal and vertical directions. Furthermore, consider the fractional difference operator $(1-L)^d$ widely used in long memory time series literature for which, by binomial expansion, we can write

$$(1-L)^d = \sum_{j=0}^{\infty} \theta_j(d) L^j,$$

where d is the order of fractional integration and

$$\begin{aligned} \theta_j(d) &= \frac{j-1-d}{j}\ \theta_{j-1}(d) = \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)}, \quad j \geq 0 \\ &\sim \frac{j^{-d-1}}{-\Gamma(d)}, \quad d \neq 0,\ \ j \to \infty. \end{aligned}$$

Similarly to fractional differences, consider also the fractional integration operator upon inversion,

$$(1-L)^{-d} = \sum_{j=0}^{\infty} \theta_j(-d) L^j$$

with

$$\theta_j(-d) = \frac{j-1+d}{j}\ \theta_{j-1}(-d) = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)}, \quad j \geq 0$$
$$\sim \frac{j^{d-1}}{\Gamma(d)}, \quad d \neq 0,\ \ j \to \infty$$

where $\Gamma(x)$ is the gamma function. Then, the Y is said to be a spatial *ARFIMA process with separable fractional difference operator* if it has the following representation

$$\begin{aligned} Y_{s,t} &= (1-L_1)^{-d_1}\,(1-L_2)^{-d_2}\,X_{s,t} \\ &= \sum_{u=0}^{\infty}\sum_{v=0}^{\infty} \theta_{u,v}(-d_1,-d_2)\,L_1^u L_2^v X_{s,t} \end{aligned} \tag{1}$$

where $\theta_{u,v}(-d_1,-d_2) = \theta_u(-d_1)\theta_v(-d_2)$ and $0 < d_k < 0.5$, $k = 1,2$, is the order of fractional integration in the horizontal and vertical directions.

Given the infinite moving average MA(∞) representation of model in (1), it follows that the spectral density of Y exists for all $\lambda_k \in (0,\pi],\ k = 1,2$ and it is equal to

$$\begin{aligned} f_y(\lambda_1,\lambda_2) &= |1-e^{-i\lambda_1}|^{-2d_1}\ |1-e^{-i\lambda_2}|^{-2d_2} f_x(\lambda_1,\lambda_2) \\ &= \left[2\sin\left(\frac{\lambda_1}{2}\right)\right]^{-d_1} \left[2\sin\left(\frac{\lambda_2}{2}\right)\right]^{-d_2} f_x(\lambda_1,\lambda_2). \end{aligned} \tag{2}$$

Since $\lim_{\lambda\to 0}\sin(\lambda)/\lambda = 1$, the behaviour of $f_y(\lambda_1,\lambda_2)$ at the origin is

$$f_y(\lambda_1,\lambda_2) \sim |\lambda_1|^{-2d_1}|\lambda_2|^{-2d_2}\ f_x(\lambda_1,\lambda_2), \qquad (\lambda_1,\lambda_2) \to (0,0). \tag{3}$$

3 See surface temperatures

As a real application, we consider composite TMI satellite images over the Pacific Ocean. Specifically, the observations represent monthly averages of sea surface temperatures (SST) on a (120×80) grid. These data were previously analyzed by [11] who used spectral methods to fit a stationary Gaussian process on a periodic lattice. The data considered here refer to March 2020. SST is known to be a significant driver of global weather and climate patterns and to play important roles in the exchanges of energy, momentum, moisture and gases between the ocean and atmosphere. As such, its knowledge is essential to understand and assess variability and long-term changes in the Earth's climate. SST is an essential parameter in weather prediction and atmospheric model simulations, and is also important for the study of marine ecosystems.

Inspired by equation (3), Figure 1 shows the values of the periodograms (as an empirical analogue of the spectral density) in log-log-coordinates obtained from the row and column series. The negative slopes suggest that an estimate of the LM parameters in the North-South and East-West directions are, respectively, equal to $\hat{d}_1 = 0.60$ and $\hat{d}_2 = 0.56$. In all cases the long memory appears quite strong for these data and this poses questions about the stationarity of the process. This makes identification of stochastic long memory even more difficult, because typical long memory features may be confounded with nonstationary components. Identifying and assessing possible long-memory components is thus essential for correct inference about the non-stationary components and one main question here is whether there is evidence for a systematic or stochastic trend in the data. This issue will be considered in an extended version of the present work.

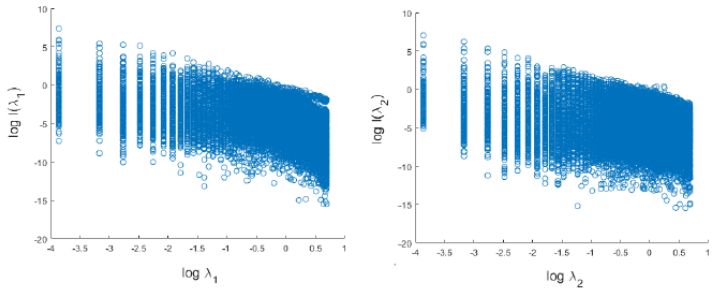


Fig. 1 *Log-periodogram versus the log-frequencies. Both figures explore the existence of possible different LM dependence in the North-South (left) and East-West (right) directions*

References

1. Lin, G. and Chen, X. and Fu, Z.: Temporal-spatial diversities of long-range correlation for relative humidity over China. Physica A: Statistical Mechanics and its Applications **383**, 2, 585–594 (2007)
2. Percival, D.B. and Rothrock, D.A. and Thorndike, A.S. and Gneiting, T.: The variance of mean sea-ice thickness: effect of long-range dependence. Journal of Geophysical Research **113**, C01004 (2008)
3. Anh, V.V. and Lunney, K.E.: Parameter Estimation of Random Fields with Long-Range Dependence. Mathematical and Computer Modelling **21**, 9, 67–77 (1995)
4. Smith, H. F.: An empirical law describing heterogeneity in the yields of agricultural crops. Journal of Agricultural Science **28**, 1–23 (1938)
5. Whittle, P.: On the variation of yield variance with plot size. Biometrika **43**, 337–343 (1956)
6. Martin, R. J.: On the design of experiments under spatial correlation. Biometrika **73**, 2, 247–277 (1986)
7. Lavancier, F.: Long memory random fields. In: Dependence in Probability and Statistics, pp. 195-220. Springer, New York (2006)

8. Martin, R. J.: A subclass of lattice processes applied to a problem in planar sampling. Biometrika **66**, 2, 209–217 (1979)
9. Boissy, Y. and Bhattacharyya, B. B. and Li, X. and Richardson, G. D.: Parameter estimates for fractional autoregressive spatial processes. The Annals of Statistics **33**, 2553–2567 (2005)
10. Beran, J. and Ghosh, S. and Schell, D.: On least squares estimation for long-memory lattice processes. Journal of Multivariate Analysis **100**, 2178–2194 (2009)
11. Fuentes, M.: Approximate likelihood for large irregularly spaced data. Journal of the American Statistical Association **102**, 321–331 (2007)

A Bayesian non parametric approach for bias correction for underreported data.

Un approccio Bayesiano non parametrico per la correzione dei dati sotto riportati.

Serena Arima, Giuseppe Pasculli and Silvia Polettini

Abstract Data quality is emerging as an essential characteristics of all data driven processes. The implications that data quality issues in computing health or vital statistics may have on government intervention policies and distribution of financial resources highlight the relevance of the problem. In this paper, we deal with the issue of underreporting, paying particular attention to its effects on the estimation of the prevalence of a phenomenon. We propose a non parametric compound Poisson model that allows for the estimation of the reporting probabilities. The proposed model will be applied to a data set concerning early neonatal mortality in Minas Gerais, Brazil. Comparisons of the estimates obtained under several alternative models reveal that the proposed approach is accurate and particularly suitable when there is no prior information about the reporting probability.

Abstract Negli ultimi anni e durante la pandemia, si é confermato il fatto che i dati di qualità sono l'elemento fondamentale per tutti i processi di decisione basati sui dati. Infatti, dati collezionati in modo viziato portano a decisioni distorte con conseguenze su azioni politiche e distribuzioni di fondi. Ció è particolarmente importante se si tratta di sanitá e salute pubblica. In questo lavoro, consideriamo il problema della distorsione dei dati con particolare interesse agli effetti che questo ha sulla stima della prevalenza di un fenomeno. Proponiamo un modello non parametrico di Poisson per la stima delle probabilitá di tale distorsione. Il modello è applicato a dati relativi alla mortalitńeonatale in Minas Gerais, Brasile. Le stime verranno confrontante con modelli alternativi e l'applicazione evidenzia che il metodo

Serena Arima
Departiment of history, society and social sciences, University of Salento, Lecce, Italy, e-mail: serena.arima@unisalento.it;
Giuseppe Pasculli
Department of Computer, Control, and Management Engineering Antonio Ruberti (DIAG), University of Rome "La Sapienza", Rome, Italy, e-mail: giuseppe.pasculli@uniroma1.it ;
Silvia Polettini
Department of Political Sciences, University of Rome "La Sapienza", Rome, Italy, e-mail: silvia.polettini@uniroma1.it

proposto é particolarmente promettente laddove non ci sono informazioni a-priori circa la qualità del dato raccolto.

Key words: Compund Poisson model, hierarchical models, MCMC, Underreporting probabilities,

1 Introduction

In order for scientists to derive reliable statistical analyses, useful for taking appropriate data driven decisions, good quality data are mandatory. As experienced in the last year, inaccurate data collection leads to inappropriate conclusions even when accurate and complex statistical methodologies had been performed. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. In underdeveloped and developing countries, the poor quality of the available data often impairs estimation of economic, health, and social indicators. An example is provided by the case study of this paper, originally presented in [4], where early neonatal mortality risk in Minas Gerais, Brazil, during 1999-2001 is studied. The state of Minas Gerais presents heterogeneous characteristics and a relevant socio-economical inequality. In this context, it is expected that official figures do not reflect the real situation with reports most likely overlooking deaths occurring soon after birth. If such phenomena are not taken into proper account, critical statistics may be underestimated, jeopardizing the concept of effective government intervention policies and the allocation of financial resources.
The problem of underreporting is well known in the literature. The bias caused by a flawed data reporting method is widely discussed in the statistical literature by using hierarchical models that accommodate truncated or censored observations. [4] propose a bias correction method based on a compound Poisson model for count data that includes an area-specific reporting probability, whose uncertainty is accounted for in the model. In the proposal, areas are clustered according to their data quality. Since underreporting probabilities might reflect socio-economic, political and/or demographic characteristics of each region, we focus on modelling the underreporting rates in different areas. Building on the idea in [4], we consider a Bayesian semi-parametric extension of the compound Poisson model: following the approach described in [8], we propose a new method to introduce covariates to define the clustering structure for the underreporting probability ε_i.

2 Modelling underreporting probability

The bias problem induced by defectively reported data is widely discussed in the statistical literature: a great variety of models that accommodate truncated or censored

observations has been developed. Several extensions of the Poisson model have been proposed for counts subject to underreporting ([1], [2], [7]). The compound Poisson model (CPM) is an alternative approach to deal with potentially underreported counts. It allows for the joint modeling of the event occurrence rates and the associated reporting probabilities. [4] discussed issues about the identifiability of the CPM and proposed an alternative Bayesian hierarchical formulation of the CPM that only requires reliable prior information about the reporting process for areas experiencing the best data quality. They rely on the fact that areas within the region under study can be clustered into homogeneous groups reflecting the data quality of each area. The clusters may be defined based on experts' opinion or applying some clustering technique to auxiliary data quality indicators provided by previous studies and surveys.

2.1 A compound Poisson model for underreporting probabilities

Consider a region consisting of m areas and denote by Y_i the observed counts in area i ($i = 1, ..., m$). Denote by T_i the number of events of interest, and assume

$$Y_i \sim Poisson(E_i\theta_i),$$

where θ_i is the relative risk and E_i is a known offset representing the expected number of events in the $i-$th area. T_i is possibily underreported, in the sense that for each record $j = 1, \ldots, T_i$ in area i the event is actually reported with probability ε_i, namely $Z_{ij} \sim Bernoulli(\varepsilon_i), i = 1 \ldots, T_i$ independently, and independent on T_i.

The observed counts are defined as $Y_i = \sum_{j=1}^{T_i} Z_{ij}$ and therefore, $Y_i|T_i, \varepsilon_i \sim Bin(T_i, \varepsilon_i)$. Combining with 2.1 and marginalising over T_i this is the so called compound Poisson model (CPM), under which

$$Y_i|\theta_i\varepsilon_i \sim Poisson(E_i\theta_i\varepsilon_i).$$

Let us further assume that we can relate the relative risks to a set of covariates $X_1, ..., X_p$:

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}.$$

As described above, the parameter ε_i defines the reporting probability in the $i-$th area: low values of ε_i indicate areas whose observed counts are underreported. [4] assume that areas can be a-priori clustered according to their data quality. Indeed, they specify

$$\varepsilon_i = 1 - h_i^T \gamma$$

where $(h_{1i}, ..., h_{Ki})^T$ is the cluster indicator defined according to the following split-coding scheme: if area i belongs to cluster j then $h_{li} = 1$ for all $l \leq j$ and $h_{li} = 0$ otherwise. They also assume that $\gamma \in [0, 1)$ and $\sum_{j=1}^{K} \gamma_j < 1$ to ensure non-null mean for the associated Poisson distribution.

To ensure identifiability of the model, informative priors about the reporting probabilities and the induced clustering structure among the areas are specified: in particular, [4] fix the number of cluster and model the probability of underreporting according to a-priori knowledge about data quality. Strongly informative priors are defined, especially for the areas that are supposed to be characterized by the best data quality.
However, as revealed by the real data application in [4], parameter estimates do sensibly vary according to the number of clusters and to cluster assignment. Besides this, it might be the case that prior information about data quality is not available. In this work, building on the same compound Poisson model, we consider an alternative approach: we specify a clustering structure for the reporting probability ε_i following a non parametric approach based on a dependent Dirichlet process, that allow the aggregating property of the DP to depend on covariates. Although such a specification is more complex from a theoretical as well as computational point of view, it significantly increases the flexibility of the model since it does not require a-priori knowledge of the number of clusters and it defines the clustering structure in a complete nonparametric, data driven way. Indeed, the clustering is induced by introducing covariates in the stick breaking construction of the Dirichlet process.

3 The proposed model

Let $\varepsilon^n = (\varepsilon_1, ..., \varepsilon_m)$ and $Z^n = (z_1, ..., z_m)$ denote, respectively, the entire vector of the reporting probabilities and the covariate Z used as predictor for ε.

A simple nonparametric model can be defined by introducing a DP model on the reporting probabilities:

$$Y_i|\theta_i\varepsilon_i \sim Poisson(E_i\theta_i\varepsilon_i).log(\theta_i) = \beta_i + \beta_1 X_{1i} + ... + \beta_p X_{pi} \quad (1)$$

$$\varepsilon_i|G \sim iid\ G \quad (2)$$

$$G|\alpha, G_0 \sim DP(\alpha, G_0) \quad (3)$$

For any measurable set B, the DP process has the well known stick-breaking representation [9]

$$G(B) = \sum_{j=1}^{\infty} w_j \delta_{\eta_j}(B)$$

where $\delta_{\eta_j}(\cdot)$ is the Dirac measure at η_j and $w_j = V_j \prod_{l<j}[1 - V_l]$ with $V_j|\alpha \overset{i.i.d.}{\sim} Beta(1, \alpha)$ that implies an infinite mixture representation of the distribution of the Y_i.

[3] propose a modification of the well known stick-breaking representation of the DP in which the weights are made dependent on covariates; this is achieved replacing the Beta random variables by normally distributed random variables transformed through the normal cdf. The resulting measure is defined as the probit-stick breaking (PSB) process, see also [9]. As described by [8], [3] allow for dependence on

covariates via the introduction of independent Gaussian processes indexed by the covariates as specified in the following formula:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left\{ \Phi(\eta_j(z)) \prod_{l<j} [1 - \Phi(\eta_l(z))] \right\} \delta_{\varepsilon_j}(\cdot)$$

where $\eta_j(z) = z'\gamma$.
As already discussed in [4], one of the main problems of the CPM is the lack of identifiability. Two main approaches can be used to obtain an identifiable model when sets of covariates, X and H, are used to model the relative risk θ and the reporting probability ε. The first one requires extra information desumed from independent validation datasets and requires non-overlapping and incorrelated X and H. The second one requires informative prior distribution for the overall mean reporting rate which is sufficient to complete the identifiability conditions. In our approach, identifiability is achieved by using two disjoint sets of covariates for estimating the risk and the reporting probability. Moreover, slight informative prior about the intercept parameter β_0 is specified ($\beta_0 \sim N(0, 10)$). Sensitivity analysis with respect to this choice reveal that the model is robust with respect prior specification.

4 Data application and results

We apply our model to data analyzed in [4]: our goal is to accurately estimate the relative risk of early neonatal mortality (ENM) in Minas Gerais State (MG), Brazil, one of the more deprived area of the country. Experts believe that the quality of infant mortality information produced in this country is usually underreported, especially in some areas located in northern and south-Eastern regions of the state. The observed counts of early neonatal deaths have been collected for the $m = 75$ areas from 1999-2001 and 2009-2011. In this work we consider data from 1999-2001. The ENM relative risk assumes a log-linear regression structure which includes local and spatial random effects. Five covariates are introduced in this regression model: the Municipal Human Development Index (MHDI), the proportion of mothers with more than twelve years of formal education (MomEduc), the proportion of children with weight at birth smaller than 2.5 Kg (LowWeight), the proportion of children who were born with some congenital anomaly (Anomaly) and the proportion of mothers who made seven or more prenatal visits during the pregnancy (Prenatal). In [4] the clustering structure is defined according to the quantiles of an adequacy index (AI) introduced by [5] as a measure of the quality of infant mortality data collected in Minas Gerais. In our approach this variable is used in the Dirichlet process for defining both the number of clusters as well as the clustering structure. Informative prior about the intercept parameter is specified according to the obtained estimates with a simple Poisson model.
Table 1 shows the posterior means of the parameters and the corresponding 95% credible intervals. As expected, the two models agree on the fact that only the co-

variate MHDI shows to be significant (likely non-zero effect) to explain the ENM risk for the period 1999–2001. The effect of the covariate MDHI is negative in both periods, indicating that the highest the MHDI, the smallest the ENM risk. According to the approach in [4], the model with $K = 8$ clusters resulted to be the best fitting one while the proposed model selects $K = 5$ clusters, showing a WAIC smaller than the one of the competing model. Also, risk estimates are substantially overlapping confirming that the ENM risks in the poorest areas (North and Northeast) are higher than the ones obtained for more developed regions of Minas Gerais.
These results reveal that the proposed approach is accurate and particularly suitable when there are no prior information about data quality.

Table 1 Parameter estimates and 95% credible intervals (in brackes) for the model proposed in [4] and the proposed model.

Parameter	Lopes et al. model $K = 8$ $WAIC = 603.091$	Proposed model $K = 5$ $WAIC = 598.360$
Intercept	1.986 (1.624; 2.348)	0.852 (0.092; 1.612)
MHDI	-3.369 (-3.991; -2.747)	-1.635 (-2.868; -0.402)
MomEduc	-0.033 (-1.263; 1.197)	1.452 (-4.469; 1.317)
LowWeight	-0.095 (-1.279; 1.089)	2.206 (-3.932; 6.836)
Anomaly	2.450 (-11.972; 16.872)	2.207 (-15.644; 20.057)
Prenatal	0.104 (-0.230; 0.528)	0.507 (-0.0311; 1.044)

References

1. Bailey, T. C., Carvalho, M. S., Lapa, T. M., Souza, W. V., and Brewer, M. J. Modeling of under-detection of cases in disease surveillance. Annals of Epidemiology **15(5)**, 335–343, (2005).
2. Caudill, B. S. and Mixon Jr., F. G. Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. Empirical Economics **20(2)**, 183–196, (1995).
3. Chung, Y. and Dunson, D. B. Nonparametric Bayes conditional distribution modeling with variable selection. Journal of the American Statistical Association **104**,1646–1660, (2009).
4. de Oliverira, G.L., Argiento, R., Loschi, R.H. Bias correction in clustered underreported data, Bayesian Analysis TBA, 1–32, (2020)
5. Franca, E., Abreu, D., Campos, D. and Rausch, M. C. Avaliacao da Qualidade da informacao sobre a mortalidade infantil em Minas Gerais: Utilizacao de uma metodologia simplificada (available in Portuguese). Revista Medica de Minas Gerais **16(1 suppl. 2)**, 15 –22, ((2006).
6. Hart,J.T. The inverse care law, Lancet **7696**, 405–412, (1971).
7. Oliveira, G. L., Loschi, R. H., and Assuncao, R. M. A random-censoring Poisson model for underreported data. Statistics in Medicine **36(30)**, 4873–4892, (2017).
8. Quintana, F., Mueller, P., Jara, A., MacEachern, S. The dependent Dirichlet process and related models. arXiv:2007.06129, (2021)
9. Sethuraman, J.A constructive definition of Dirichlet prior. Statistica Sinica **2**, 639–650, (1994).

Assessment of the impact of anthropic pressures on the Giglio island meadow of Posidonia oceanica

Valutazione dell'impatto delle pressioni antropiche sulla prateria di Posidonia oceanica dell'isola del Giglio

Giovanna Jona Lasinio, Gianluca Mastrantonio, Alessio Pollice, Daniele Ventura, Gianluca Mancini, Giandomenico Ardizzone[1]

Abstract We present a Bayesian Beta regression model for the assessment of anthropic pressures on the Posidonia meadows along the Giglio island coasts. The evolution of the meadows was assessed by analysis of aerial photos taken from 1968 until 2013.
Abstract *Il lavoro presenta un modello di regressione Beta bayesiano per la valutazione dell'impatto antropico sulle praterie di Posidonia presenti lungo la costa dell'isola del Giglio. L'evoluzione delle praterie è studiata attraverso l'analisi di foto aeree scattate annualmente tra il 1968 e il 2013.*

Key words: Beta regression, Posidonia oceanica, Bayesian statistics.

1. Introduction

Posidonia oceanica is the most important and widespread endemic seagrass species in the Mediterranean Sea, capable of developing large meadows from the sea surface level up to 40-45 meters depth (Duarte, 1991). It forms one of the most valuable coastal ecosystems on Earth in terms of goods and services for its ecological, physical, economic, and bio-indicator role (Vassallo et al., 2013).
Due to its wide distribution and its unique features, *P. oceanica* is protected by EU legislations and local measures both at species and at habitat levels. Even though the P. *oceanica* is protected by a legal framework its meadows are rapidly declining during the last century, mainly due to human activities, climate changes, and alien species invasion (Telesca et al., 2015).
Effective coastal zone management plans and conservation efforts on *P. oceanica* could benefit from a more profound knowledge of seagrass spatial distribution. Marine spatial planning and integrated coastal zone management are pivotal in promoting sustainable growth of maritime and coastal activities and using coastal and marine resources

[1]Giovanna Jona Lasinio, DSS - Università di Roma La Sapienza; email: giovanna.jonalasinio@uniroma1.it
Gianluca Mastrantonio, DISMA - Politecnico di Torino
Alessio Pollice, DiEF - Università di Bari Aldo Moro
Daniele Ventura, Gianluca Mancini, Giandomenico Ardizzone, DBA - Università di Roma La Sapienza

sustainably, as also recently highlighted by the European Commission (Schaefer and Barale, 2011).
Coastal benthic habitats, such as *P. oceanica*, can be described through spatial representations of discrete seabed areas associated with particular species, communities, or co-occurrences (Papakonstantinou et al., 2020), known as benthic or bionomic maps. These maps provide baseline information for research activities and maritime activities in coastal areas.
Motivated by the above, in this work we analyze human impacts on the *P.oceanica's* meadow of the Giglio island in the period 1968-2013. The main source of information is the percentage of Posidonia coverage on an area extrapolated from aerial photos. Proportional data, in which response variables are expressed as percentages or fractions of a whole, are analysed in many fields. The scale-independence of proportions makes them appropriate to analyse many biological phenomena, but statistical analyses are not straightforward. Transformations to overcome these problems are often applied, but can lead to biased estimates and difficulties in interpretation. Beta regression overcomes some problems inherent in applying classic statistical approaches to proportional data.

2. Study area and human activities

The study area is represented by the Island of Giglio (Central Tyrrhenian Sea, Italy), one of the seven primary islets, plus several smaller, composing the Tuscany Archipelago National Park (TANP). The aquatic environment of Giglio Island is characterized by the presence of a vast and almost continuous *P. oceanica* meadow thriving on matte, sand, and rock from few centimeters below the sea surface up to 37 meters depth on a gently sloping seabed. The meadow runs all around the island except for the west-south quadrant characterized by vertical cliffs and steep bottoms, a harsh environment for *P. oceanica* thriving. The upper and lower edges (*i.e.*, the landward and seaward boundaries defining the meadow) are localized at different depths and distances from the coastline. They follow the seabed slope, the hydrodynamic forces, the photosynthetic process, and the anthropogenic pressures (Montefalcone et al., 2010). The coastal area is divided into 13 zones around the perimeter of the Island (Fig.1) according to the seabed morphologies. The latter determined the visibility, which allowed the identification of Posidonia meadow limits from aerial images. Only shallow coastal areas (up to 12 meters depth) were selected for polygon editing in GIS software aimed at defining the extension of the Posidonia meadow. The same zones were divided into *Shallow* and *Deep.* No active protection is undergone on the meadow all over the area. For this reason, *P. oceanica* has been directly and indirectly threatened by several anthropic pressures such as the i) pleasure boats anchoring, ii) constructions (harbors, public works, urban and rural areas development) and agricultural practices, and iii) mining.

Assessment of the impact of anthropic pressures on the Giglio island meadow of *Posidonia oceanica*

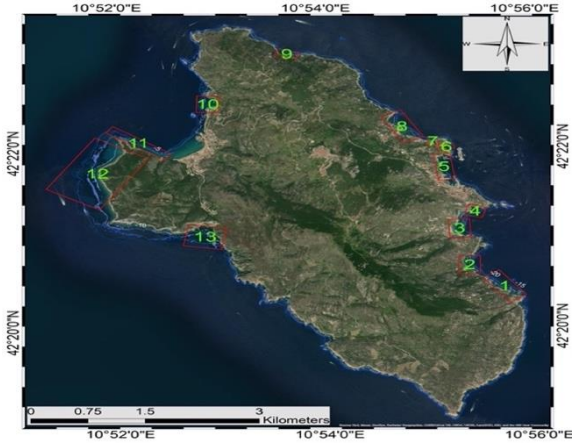


Figure 1 Study area with the 13 zones highlighted.

Anchoring: Due to the land proximity and its sheltered bays, Giglio Island represented a popular seaside destination for touristic boating, which anchoring was localized on the *P. oceanica* meadow close to the coastline. The anchoring is defined as the short-term deployment of a physical device to hold fast to the substrate by a vessel. It has been proved to disturb *P. oceanica* meadows at different levels (Deter et al., 2017;).

Constructions and agricultural practices: During the last fifty years, the island faced a massive anthropic outbreak in terms of touristic frequentation, leading to increased public works and urban and rural areas development. Coastal constructions involved **harbor enlargement** and the desalination system development. Many constructions were next to the coastline. Due to the temperate weather and the fertile soil, Giglio Island was characterized by grapevines (*Vitis vinifera*) and olive trees (*Olea europaea*) cultivations. To face the mountainous environment, terracing was adopted as agricultural practice all over the island. Terraces were built by constructing dry-stone walls, named 'greppe,' using granite blocks; landward, regolite soil was laid over the bedrock as a substrate for cultivation. Today, few terraces are actively cultivated and maintained, whereas the ones abandoned are deteriorating and collapsing, leading to landslide events and contributing to water runoff and sediment generation moving to the seaside.

Mining: Since the Roman age, mining activities have interested the island with granite, limestone, and gypsum extraction and, more recently, pyrite and further iron minerals exploitation. Granite caves, mainly localized in the eastern side of the Island from Arenella to Caldane Bays, provided monzogranite rocks up to 1950, serving all central Italy. Metamorphic and sedimentary rocks mining interested the northwestern side of the island, in the Frengo Promontory (next to Campese Bay), up to the 1960s. Caves produced limestone and gypsum, whereas pyrite and iron minerals were obtained from the Frengo mine, which closed in 1976. To move the pyrite from the mine to the barges moored in Campese Bay, a cableway was mounted on three pillars built over the *P. oceanica* meadow at 5 meters depth. Mining activities led to debris production and

dump areas, resulting in a high quantity of the reduced size of rocks, from a few centimeters up to one meter.
Each impact is recorded as intensity, presence/absence and distance between the zone and the impact source.

2.1 Further available data

Together with the above described human-activities variables, mean depth and mean slope for each zone, errors in the aerial photos, resolution of the photo, sea state when the photo was taken, are available for modelling.

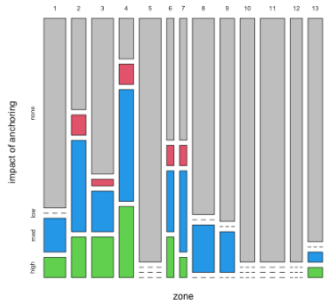
3. The model

To understand the relationship of *P. oceanica* coverage of the Giglio island coastal area with the measured impacts and environmental conditions, we used a Beta regression model(Ferraro, Cribari-Neto, 2004), that is a generalized linear model based on the Beta distribution $Y_{it} \sim Beta(\mu_{it}, \tau_{it})$ where Y_{it} is the i-th observation of Posidonia coverage at time *t*, μ_{it} is the mean of the distribution and τ_{it} the precision. Further $\text{logit}(\mu_{it}) = \beta_{0\mu}^{z_{k_i}} + \sum_{h=1}^{p} x_{hti}\beta_{h\mu}$ and $\log(\tau_{it}) = \beta_{0\tau} + \sum_{l=1}^{k} x_{lti}\,\beta_{l\tau}$ were $\{x_{it}\}$ *is the set of* available data and z_{k_i} denotes cluster membership of the zone k_i (k_i=1,…,13) where observation *i* occurs. Hence, by the same model we investigate both the influence of anthropic impacts on the Posidonia coverage and the presence of homogeneous clusters of zones. We perform the model estimation in the Bayesian setting, implementing our code in JAGS. The set of parameters' priors distributions are: for all $\beta_\mu, \beta_\tau \sim N(0{,}1000)$, $z_j \sim multinom(\pi), j = 1, \dots, 13$ and $\pi \sim Dirichlet(\alpha = 1)$. We run the MCMC sampler for 160000 iterations with a burn-in of 80000, keeping 5000 samples for inference after thinning.

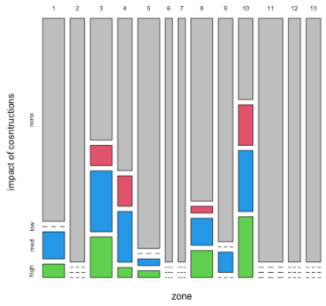
4. Exploratory data analysis and preliminary results

In Fig. 2 the mosaicplots of zones and impacts intensities are shown. It appears that only few zones are affected by multiple impacts. However, no zone is free from impacts of some kind. From further explorations of time trends (Fig.3) it appears that for some zones a decrease in the Posidonia coverage is suspected.
Preliminary results from model selection based on the DIC criterion, suggest that the distance from the impact source is not influential, while the presence/absence of the impact is important in terms of model fitting. There is evidence of 4 different groups. In Fig. 2 the grouping of zone coefficients is well described. Impacts of harbor, anchorage and mining activities are relevant factors. We are currently refining the model, in particular by adding more specific variables to the description of the precision parameter τ.

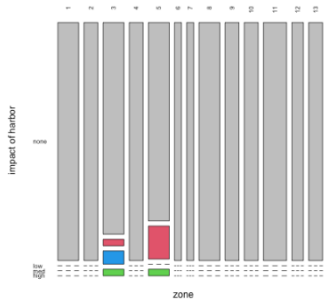
Assessment of the impact of anthropic pressures on the Giglio island meadow of *Posidonia oceanica*



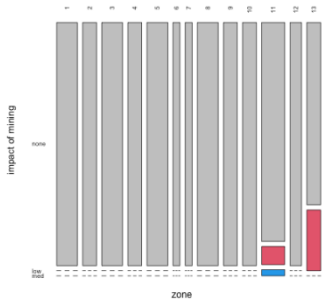
(a) anchoring



(b) constructions and agriculture



(c) harbor



(d) mining

Figure 2 *Distribution of impact levels by zone (grey = no impact, blue = low impact, green = moderate impact, red =high impact).*

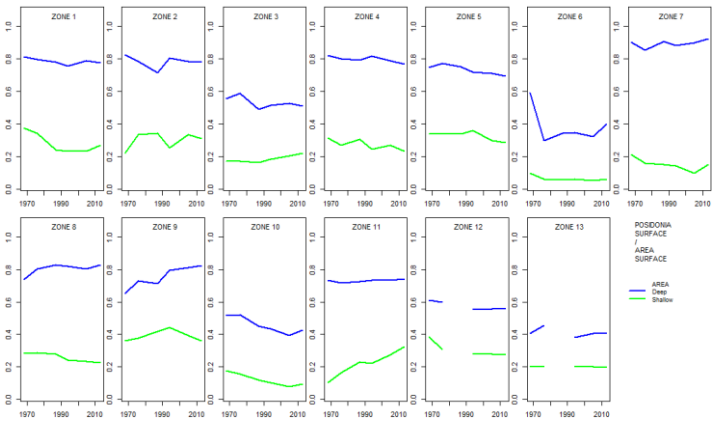


Figure 3 *Posidonia o. meadows coverage: time trend by zone and depth.*

References

1 Deter, J., Lozupone, X., Inacio, A., Boissery, P., Holon, F.: Boat anchoring pressure on coastal seabed: Quantification and bias estimation using AIS data. Mar. Pollut. Bull. **123**, 175–181 (2017)

2 Douma, J.C., Weedon, J.T.: Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods Ecol Evol. **10**, 1412– 1430 (2019)

3 Duarte, C.M.: Seagrass depth limits. Aquat. Bot. **40**, 363–377 (1991)

4 Ferrari, S.L.P., Cribari-Neto, F.: Beta Regression for Modeling Rates and Proportions. Journal of Applied Statistics **31**(7), 799–815 (2004)

5 Montefalcone, M., Parravicini, V., Vacchi, M., Albertelli, G., Ferrari, M., Morri, C., Bianchi, C.N.: Human influence on seagrass habitat fragmentation in NW Mediterranean Sea. Estuar. Coast. Shelf Sci. **86**, 292–298 (2010)

6 Papakonstantinou, A., Stamati, C., Topouzelis, K.: Comparison of true-color and multispectral unmanned aerial systems imagery for marine habitat mapping using object-based image analysis. Remote Sens. **12** (2020)

7 Schaefer, N., Barale, V.: Maritime spatial planning: Opportunities & challenges in the framework of the EU integrated maritime policy. J. Coast. Conserv. **15**, 237–245 (2011)

8 Telesca, L., Belluscio, A., Criscoli, A., Ardizzone, G., Apostolaki, E.T., Fraschetti, S., Gristina, M., Knittweis, L., Martin, C.S., Pergent, G., Alagna, A., Badalamenti, F., Garofalo, G., Gerakaris, V., Louise Pace, M., Pergent-Martini, C., Salomidi, M.: Seagrass meadows (Posidonia oceanica) distribution and trajectories of change. Sci. Rep. **5**, 1–14 (2015)

9 Vassallo, P., Paoli, C., Rovere, A., Montefalcone, M., Morri, C., Bianchi, C.N.: The value of the seagrass Posidonia oceanica: a natural capital assessment. Mar. Pollut. Bull. **75**, 157–167 (2013)

10 Waycott, M., Duarte, C.M., Carruthers, T.J.B., Orth, R.J., Dennison, W.C., Olyarnik, S., Calladine, A., Fourqurean, J.W., Heck, K.L., Hughes, A.R., Kendrick, G.A., Kenworthy, W.J., Short, F.T., Williams, S.L.: Accelerating loss of seagrasses across the globe threatens coastal ecosystems. Proc. Natl. Acad. Sci. U. S. A. **106**, 12377–12381 (2009)

Session of solicited contributes SS17 – *Functional Data Analysis Methodologies for Quality Assessment*
Organizer and Chair: Elvira Romano

Improving the quality of questionnaires via the combined use of functional outlier detection and Item Response Theory

Perfezionare la qualità dei questionari attraverso l'uso combinato dell'analisi dei dati funzionali e dell'Item Response Theory

Francesca Fortuna, Fabrizio Maturo, and Tonio Di Battista

Abstract The use of questionnaires for evaluation is widely accepted in many areas; thus the assessment of their quality plays an essential role. Understanding how to improve a test by eliminating items that are wrong or can be misinterpreted requires instruments that adapt to specific contexts. This article proposes a combined use of functional data analysis and Item Response Theory as a tool for identifying outliers in questionnaires. The basic idea is that item characteristics curves of a questionnaire can be treated as functional data and, in the context of performance evaluation, possible outliers should be evaluated for different competitors' skill levels. The final aim of the paper is to suggest a methodology for assessing outliers for improving questionnaires' quality in different scenarios.

Abstract *L'uso dei questionari per la valutazione è ampiamente diffuso in molte aree e la valutazione della qualità degli stessi gioca un ruolo essenziale. Capire come migliorare un test eliminando gli item sbagliati o che possono essere mal interpretati, richiede strumenti che si adattino a diversi contesti. Questo articolo propone l'uso combinato dell'analisi dei dati funzionali e dell'Item Response Theory per identificare possibili outliers nei questionari. L'idea di base è che le curve caratteristiche degli item possono essere trattate come dati funzionali e, nel contesto della valutazione delle performance, possibili outliers dovrebbero essere considerati per diversi livelli di abilità dei concorrenti. L'obiettivo finale dell'articolo è quello di suggerire una metodologia di valutazione degli outliers per migliorare la qualità dei questionari in diversi scenari.*

Key words: FDA, Functional Outlier Detection, IRT, ICC, Questionnaire Quality

Francesca Fortuna
Roma Tre University, Rome, Italy. e-mail: francesca.fortuna@uniroma3.it

Fabrizio Maturo
University of Campania Luigi Vanvitelli, Caserta, Italy. e-mail: fabrizio.maturo@unicampania.it

Tonio Di Battista
G. D'Annunzio University of Chieti-Pescara, Italy. e-mail: tonio.dibattista@unich.it

1 Introduction

The use of questionnaires and tests for evaluation is widely accepted in many areas, e.g. entrance tests for degree courses, university exams, and competitions between students in various sectors. Therefore, the evaluation of questionnaires' quality plays an essential role both for the judgment of participants and the credibility of people providing the test. Nowadays, the consequences of errors in questionnaires are dangerous both from a judicial and substantive point of view. Understanding how to improve a test by eliminating items that are wrong, or that participants can misinterpret, requires methodological tools able to capture the peculiarities of different contexts.

Item Response Theory (IRT) models are widely used for the analysis of test data as the former describe the relationship between the response behaviour of subjects to a set of items and the underlying latent trait, which is indirectly measured by the items [8]. This relation is given by the Item Characteristic Curve (ICC), which shows how the probability of success on a test item changes as the level of the latent trait varies. Classical IRT models assume that the ICC is included in a restricted class of functions, defined by specific mathematical models, such as the logistic function [8, 1]. Parameters' estimate of IRT models can become significantly biased in the presence of outliers. However, the practice of removing outliers is controversial, as it may result in data information loss. Indeed, outliers' detection can provide valuable insights into subjects, items, or tests [11]. For example, outliers may be a result of aberrant item responses by persons, and/or differentially functioning test items [6]. For these reasons, outliers detection is a key issue in questionnaires' quality evaluation. Classical outlier detection techniques, such as box-plots, determine outliers from total scores, and thus they are not able to identify atypical patterns of responses across the items. Indeed, it may occur that participants with similar total scores present very different responses. To solve this issue, person fit statistics generated from a variety of different approaches including item residual, least squares, and IRT methods have been proposed in the literature [10]. Therefore, outlying responses, items, and/or individuals can be removed obtaining more robust estimates of IRT models' parameters [4].

An original approach to deal with ICCs using Functional Data Analysis (FDA) has been proposed by some recent research [5, 3, 12, 14]. Effectively, the FDA approach can provide many advantages also in the educational context. First, it allows catching specific characteristics of ICCs without forcing them into a predefined mathematical model. Second, it enables the evaluation of items shape for all the levels of the latent trait. Finally, the graphical inspection of ICCs might show problematic items, which should be revised or excluded from the test [9]. Starting from the fundamental idea of the above-mentioned previous literature, this article proposes an original methodological framework based on the combined use of FDA and IRT to identify outliers in questionnaires. In addition, this study provides a local functional outliers' detection approach to take into account different skill levels of competitors in the context of IRT. The final aim of this study is to suggest a methodology for improving questionnaires' quality in different scenarios.

2 FDA in the context of IRT

The basic representation of IRT models for dichotomous items [8] is given by:

$$P(X = 1|\boldsymbol{\theta}) = f(\boldsymbol{\eta}, \boldsymbol{\theta}), \tag{1}$$

where X represents the binary response on the test item, with $X = 1$ representing the correct one; $\boldsymbol{\eta}$ is a vector of parameters, which denotes the characteristics of the item; $\boldsymbol{\theta}$ represents a proficiency parameter for each subject; and f is a function which defines the relationship among the item parameters, the subject ability and the probability of a correct response [9]. The most common choice for f in (1) is the logistic function, which leads to the so-called one (1PL), two (2PL), and three (3PL) parametric logistic models. The latter model [1] provides the general form and specifies the *ICC* of the i-th item ($i = 1, \cdots, n$) as follows:

$$ICC_i(\theta) = P(X_{ji} = 1|\theta_j, \gamma_i, \alpha_i, \beta_i) = \gamma_i + (1 - \gamma_i)\frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]}, \tag{2}$$

where X_{ji} denotes the observed response of the j-th subject ($j = 1, ..., J$) to the i-th item ($i = 1, .., n$), θ_j indicates the ability of the j-th subject, while γ_i, α_i and β_i are the pseudo-guessing, the discrimination, and the difficulty parameter of the i-th item, respectively. For a detailed explanation of the item parameters, see [8, 1]. In this research, the 2PL model is considered as a special case of (2), i.e. by fixing $\gamma_i = 0$ [1]. Following the FDA approach [13], the data consists of n ICC functions, $ICC_1(\theta), ..., ICC_n(\theta)$, and each of them can be expressed as a linear combination of a given number of basis functions, $\phi_k(\theta)$, $k = 1, ..., K$, as follows:

$$ICC_i(\theta) = \sum_{k=1}^{K} c_{ik}\phi_k(\theta), \tag{3}$$

where c_{ik} is the coefficient of the i-th item and k-th basis function $\phi_k(\theta)$.

3 Local and global functional outliers detection of ICCs

Identifying possible outliers is essential during the exploratory analysis because they can significantly bias any statistical analysis. To discover outlying data in the functional domain, many statistical depth measures have been extended to the FDA context to assess the centrality of functional data with respect to the sample [2]. Thus, functional observations with large depth are close to the centre of the sample whereas low depth observations are candidates to be appointed as outliers.
In this article, we extend the Band Depth (BD) approach [7] to ICCs curves by focusing on a local and global functional outlier detection procedure. BD is based on the graphical representation of the functions and makes use of the bands de-

fined by their graphs on the plane. The graph of an ICC is the subset of the plane $G(ICC) = \{(\theta, ICC(\theta)) : \theta \in \Theta\}$. The sample band depth of the i-th $ICC(\theta)$ is:

$$BD_{i,n,R}\Big(ICC_i(\theta)\Big) = \sum_{r=2}^{R} BD_{i,n}^{(r)}\Big(ICC_i(\theta)\Big), \tag{4}$$

where R is the number of curves determining a band, with $2 \leq R \leq n$, and

$$BD_{i,n}^{(r)}\Big(ICC_i(\theta)\Big) = \frac{\sum_{1 \leq i_1 < i_2 < .... < i_r \leq n} I\Big\{G(ICC(\theta) \subseteq B(ICC_{i1}(\theta), ..., ICC_{ir}(\theta))\Big\}}{\binom{n}{r}}, \tag{5}$$

represents the fraction of bands determined by r different sample curves containing the whole graph of the curve $ICC_i(\theta)$, and $I(\cdot)$ denotes an indicator function with value 1 if $ICC_i(\theta)$ is within the band and 0 otherwise. The larger the value of $BD_{i,n,R}\Big(ICC_i(\theta)\Big)$, the more central the curve is. The global outlier detection procedure is based on the computation of $BD_{i,n,R}\Big(ICC_i(\theta)\Big)$, for each function, over the whole ability domain. Instead, the local outlier detection method starts from the basic idea that the search for outliers should take into account a specific ability interval. To this purpose, the domain can be divided into "windows", each identifying a specific skill level. The band depth of a specific window can be defined as follows:

$$BD_{i,n,R}^{[w]}\Big(ICC_i(\theta_w)\Big) = \sum_{r=2}^{R} BD_{i,n}^{[w](r)}\Big(ICC_i(\theta_w)\Big), \tag{6}$$

where $w = 1, 2, ..., W$, remarks we are considering the band depth for the ability subinterval w-th. The choice of W, and also the particular subinterval(s) to take into account in the local outliers' assessment phase, should be done according to the aim of the questionnaire. A possible empirical solution could be, for example, to fix $W = 3$ and thus split the ability into low, middle, and high skill levels, respectively. In other words, the so-called "local functional BD procedure" aims to discover outliers based on the specific target of the test. According to this procedure, it is reasonable to build local rankings and a functional box-plot for each window leading to different outliers.

4 Conclusions

Outliers detection in the context of IRT is essential for improving the quality of a questionnaire. To this end, the use of the FDA approach yields several advantages because ICC curves can be considered as functional data and functional tools can be exploited to get additional insights. The search for global outliers could be few interesting because, the test could have a specific target, e.g. evaluate the best participants as in the case of competitions (the focus is on high skill levels) or evaluate

those who are sufficient as in the case of suitability exams (the focus is on medium skill levels), etc. In the above circumstances, a global outliers' detection procedure could lead to eliminate items due to an outlying behaviour in parts of the ability domain that are few interesting for the purpose of the test. Instead, looking for local outliers makes sense according to the specific target of the questionnaire. Therefore, the final aim of this study is to propose a combined use of IRT, FDA, and functional outliers' detection via a local band depth approach.

References

1. Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F., Novick, M. (eds.) Statistical Theories of Mental Test Scores, pp. 397—479. Addison-Wesle, Boston (1968)
2. Cuevas, A., Febrero, M., Fraiman, R.: Robust estimation and classification for functional data via projection-based depth notions. Comput. **22**, 481—496 (2007)
3. Di Battista, T., Fortuna, F.: Clustering dichotomously scored items through functional data analysis. Electron. J. Appl. Stat. Anal. **9**, 433—450 (2016)
4. Felt, J. M., Castaneda, R., Tiemensma. J., Depaoli, S.: Using Person Fit Statistics to Detect Outliers in Survey Research. Front Psychol. **26**, doi: 10.3389/fpsyg.2017.00863 (2017)
5. Fortuna, F., Maturo, F.: K-means clustering of item characteristic curves and item information curves via functional principal component analysis. Qual. Quant. **53**, 2291-–2304 (2019)
6. Karabatsos, G.: Comparing the aberrant response detection performance of thirty-six person fit statistics. Appl. Meas. Educ. **16**, 277—298 (2003)
7. Lopez-Pintado, S., Romo, J.: On the Concept of Depth for Functional Data. J. Am. Stat. Assoc., **104**, 718—734 (2019)
8. Lord, F., Novick, M.: Statistical theories of mental test scores (with contributions by A. Birnbaum). Addison-Wesley, Reading (1968)
9. 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)
10. Meijer, R.: Outlier detection in high-stakes certification testing. J. Educ. Meas. **39**, 219—233 (2002)
11. Orr, J., Sackett, P., DuBois, C.: Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. Pers. Psychol. **44**, 473—486 (1991)
12. Ramsay, J.: Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika **56**, 611—630 (1991)
13. Ramsay, J. O., Silverman, B. W.: Functional data analysis (2nd ed.). New York: Springer (2005)
14. Rossi, N., Wang, X., Ramsay, J.: Nonparametric item response function estimates with the em algorithm. J. Educ. Behav. Stat. **27**, 291—317 (2002)

Assessing government effectiveness over time: a functional data analysis approach

Valutare l'efficacia del governo nel tempo: un approccio di analisi funzionale dei dati

Alessia Naccarato, Francesca Fortuna and Silvia Terzi

Abstract The paper deals with the study of the evolution of a service quality indicator concerning institutional quality; in particular government effectiveness. In this context, the aim is to rank countries according to both the level and the temporal dynamic of their government effectiveness. To this end, a government effectiveness index is analysed by means of a functional data analysis approach, which considers the index as a function of time, so that it is possible to study the complete behaviour of the trajectory. Resorting to a functional tool, called area under the curve, an overall ranking is obtained, that is an ordering reflecting the behaviour of the functions across the entire domain.

Abstract *L'articolo intende studiare l'evoluzione di un indicatore di qualitá dei servizi delle istituzioni; in particolare dell'efficacia di governo. L'obiettivo é quello di classificare i paesi secondo il livello e la dinamica temporale dell'efficacia di governo. A tal fine, un indice di efficacia governativa é analizzato attraverso l'approccio dell'analisi funzionale che considera l'indice come una funzione del tempo, permettendo di studiarne l'andamento. Sfruttando uno strumento dell'analisi funzionale, chiamato area sotto la curva, si ottiene un ordinamento globale, cioé un ordinamento capace di riflettere il comportamento delle funzioni in tutto il dominio.*

Key words: Governance effectiveness, Functional data, Area under the curve, Overall rank

Alessia Naccarato
Roma Tre University, Rome, e-mail: alessia.naccarato@uniroma3.it

Francesca Fortuna
Roma Tre University, Rome, e-mail: francesca.fortuna@uniroma3.it

Silvia Terzi
Roma Tre University, Rome, e-mail: silvia.terzi@uniroma3.it

1 Introduction

Although the concept of institutional quality or governance is widely discussed among politicians and scholars, there is not yet a strong consensus around an established definition. It certainly includes the ability of government to effectively formulate and implement sound policies, which in turn is related to government effectiveness [8]. The definition of these multidimensional latent concepts and the identification of the variables or elementary indicators on which a composite institutional quality indicator can be based on, is well beyond the scope of our paper. Instead, we will resort to the Government Effectiveness (GE) index calculated by the World-Wide Governance Indicators (WGI) Project for over 200 countries since 1996. The GE index captures perceptions of the quality of public services, the quality of public administration and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to these policies.
Analysing government effectiveness over time is a key issue in assessing country improvement and in disseminating and encouraging good practices. For this purpose, countries are usually ranked according to the GE index values, thus summary measures, such as absolute or percentage changes in rank order in different time periods, are used to analyse the evolutionary trend. However, these measures may lead to paradoxical conclusions. For example, when a country starts from a low government effectiveness, even a slight increase can lead to high rates of change and vice versa. Thus, the assessment of changes in government effectiveness over time and across countries is an open issue due to the difficulty of establishing a unique ranking among countries and periods.
To provide a solution to this problem, we propose to analyse the GE index through the functional data analysis (FDA) approach [9, 4]. In this context, the GE index is considered as a function of time and it is possible to study the complete behaviour of the trajectory instead of working with the vector of discrete observations in different time periods [5, 6]. In particular, we propose the use of a functional instrument, the area under the curve, which allows to order the functions, providing the so-called overall ranking. It is important to stress that the area under the curve accounts for the trend of the function in the entire domain, thus highlighting both its level and its evolution over time.
The paper is organized as follows: Section 2 introduces the area under the curve for the overall ranking; Section 3 shows an application of the suggested method on the GE index for 27 European countries and Section 4 presents some concluding remarks.

2 Cross-countries rankings over time using the FDA approach

The FDA approach refers to the analysis of curves in a continuous domain and assumes the existence of unknown smooth functions, which generate and underlie the

data [9, 4]. Since in real applications sample curves are observed at discrete sampling points, $t_l \in \mathscr{T}$, with $l = 1,2,..,L$ and $\mathscr{T}$ a real interval over which data are collected, the raw data require a preliminary treatment before applying the FDA techniques. In particular, when the discrete data are noisy measurement of the trajectories, the functional form is reconstructed using basis expansion methods. Then, for each i-th observation, $i = 1,2,...,n$, the smooth function, $y_i(t)$, is expressed as a linear combination of a given number of basis functions, $\phi_k(t)$, $k = 1,...,K$, and basis coefficients, a_{ik}, as follows:

$$y_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t). \tag{1}$$

Different basis systems can be adopted. A common choice is to use B-splines basis functions due to their flexibility and useful mathematical properties [1].
The functional representation of governance effectiveness indicators yields valuable insights into the time dynamics of the phenomenon. Specifically, it is possible to provide a unique ranking among the countries and time periods with the aid of the area under the curve, a scalar measure that returns a single ordering of the functions for the whole time span. Starting from a sample of n governance effectiveness functions, $y_1(t), y_2(t), ..., y_n(t)$, countries can be sorted in descending order according to the area under the curve, say A_i [2]:

$$A_i = \int_{\mathscr{T}} y_i(t)\,dt. \tag{2}$$

Clearly, the greater the area under the curve, the greater the government effectiveness in the entire domain. The empirical distribution of the area values can be used to define groups of countries with different levels of governance effectiveness by establishing cutoffs, such as quartiles. At the same time, if we are interested in defining a local ordering, we can resort to a truncated version of the area under the curve by defining the integral in (2) for distinct intervals of the domain.

3 Application

The data used in this paper refer to the GE index of the 27 European countries over the period 1996 to 2020 [8]. The GE is obtained by aggregating 45 basic indicators (variables) from a large number of different sources. Some of these variables are: quality of bureaucracy, quality of road infrastructure, quality of primary education, satisfaction with public transportation system, quality of health care system, allocation and management of public resources for rural development, infrastructure disruption, state failure, policy instability. The 45 variables are aggregated by the WGI Project using a statistical tool known as an unobserved components model (UCM) [7, 3].

The GE time series of each country is converted into a function adopting a B-splines basis expansion as in (1), with $K = 5$ cubic B-splines basis, chosen by cross validation. Then, the area in (2) is computed for each country providing an overall ranking of the functional observations. Specifically, Finland, Denmark, Sweden, Netherlands and Luxembourg are at the top of the overall ranking; while Poland, Italy, Croatia, Bulgaria and especially Romania are in the lowest positions.

The countries are classified into four GE groups according to the quartiles of the

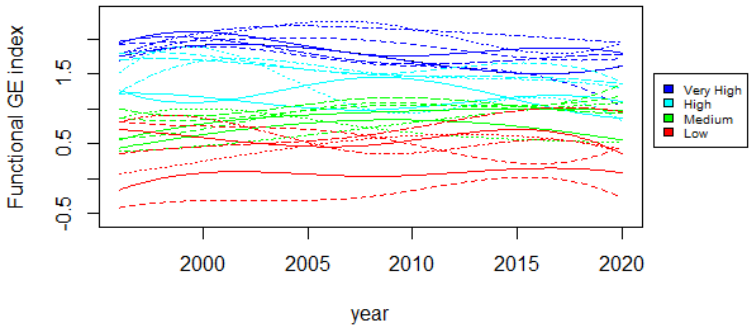


Fig. 1 Functional GE index for the 27 European countries according to their GE levels.

empirical distribution of the area under the curve. Fig. 1 shows the reconstructed functional GE index for the 27 European countries, by classifying them according to their GE level: very high GE (in blue), high GE (in cyan), medium GE (in green) and low GE (in red). The groups are balanced: they all consist of 7 countries, except for the medium group, which consists of 6 countries.

4 Conclusions

An index which captures perceptions of the quality of public services moves slowly over time, because the effects of public policies are felt in the long term. For this reason, the evaluation of the index in a single moment, however useful, cannot reflect the governmental effectiveness of a country. In this context, the FDA approach proves to be a useful tool for the evaluation of a service quality indicator. Specifically, one of the main advantages of the functional approach is to provide an overall ranking able to take into account the time dynamics of the index with a single scalar measure.

References

1. De Boor, C.: A practical guide to splines (revised ed.). Springer, New York (2001)
2. Di Battista, T., Fortuna, F., Maturo, F.: BioFTF: An R package for biodiversity assessment with the functional data analysis approach. Ecolind **73**, 726–732 (2017)
3. Bradley, E., Morris, C.: Limiting the Risk of Bayes and Empirical Bayes Estimators – Part 1: The Bayes Case. J. Am. Stat. Assoc. **66**, 807–815 (1971)
4. Ferraty F., Vieu P.: Nonparametric functional data analysis. Springer, New York (2006)
5. Fortuna, F., Naccarato, A., Terzi, S.: Building composite indicators in the functional domain: a suggestion for an evolutionary HDI. In: Perna, C., Salvati, N., Schirripa Spagnolo, F. (eds.) Book of Short Papers SIS 2021, pp. 1045—1050. Pearson (2021)
6. Fortuna, F., Naccarato, A., Terzi, S.: Functional cluster analysis of HDI evolution in European countries. In: Porzio, G.,C., Rampichini, C., Bocci, C. (eds.) CLADAG 2021, Book of Abstracs and Short Papers, 13th Scientific Meeting of the Classification and Data Analysis Group, Firenze, September 9–11, 2021, pp. 336-339. Firenze University Press (2021)
7. Goldberger, A.: Maximum Likelihood Estimation of Regressions Containing Unobservable Independent Variables. International Economic Review **13**, 13—15 (1972)
8. Kaufmann, D., Kraay, A., Mastruzzi, M.: The Worldwide Governance Indicators: Methodology and Analytical Issues. World Bank Policy Research Working Paper No. 5430 (September 2010), Available via DIALOG.
https://ssrn.com/abstract=1682130
9. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edition. Springer, New York (2005)

Mining Distributed Acoustic Sensing data for vehicle traffic monitoring

Antonio Balzanella and Rosanna Verde

Abstract In recent years Distributed Acoustic Sensing based on optic fibers is receiving a lot of attention because of the possibility to detect space-time data over extended spatial regions. In this paper we focus on using DAS data for monitoring vehicle traffic and for detecting vehicle typology. We test the results on real data coming from an experiment which uses an optic fiber installed parallel to a road.
Abstract *Negli ultimi anni i sensori acustici distribuiti basati su fibra ottica stanno assumendo un ruolo rilevante in numerosi campi applicativi poichè consentono di misurare vibrazioni su superfici molto estese. In questo articolo ci focalizziamo sull'uso dei DAS per il monitoraggio del traffico veicolare e per l'individuazione della tipologia dei veicoli. La strategia proposta è valutata su dati reali riguardanti una fibra ottica disposta parallelamente ad una strada aperta al traffico veicolare.*

Key words: Distributed Acoustic Sensing, clustering, 2D-wavelet.

1 Introduction

As the number of vehicles has increased significantly, traffic congestion has become a daily problem for each of us. This motivates huge investments of transportation agencies in developing vehicle traffic monitoring capable of collecting data such as the number of vehicles, types of vehicles, and vehicle speed. Following the taxonomy provided in [5], vehicle classification systems can be classified into three categories: in-road-based, over-road-based, and side-road-based. In-road-based systems use sensors installed on or under the pavement of a roadway, such as piezoelectric

Antonio Balzanella
Università della Campania "Luigi Vanvitelli", e-mail: antonio.balzanella@unicampania.it

Rosanna Verde
Università della Campania "Luigi Vanvitelli" e-mail: rosanna.verde@unicampania.it

sensors, magnetometers, vibration sensors. The main drawback of these systems is the high installation cost. Over-road-based systems use sensors, such as cameras, installed over the roadway. The main drawback of these systems is the sensitivity of their performances to weather and lighting conditions. Another important problem is the driver privacy protection. Finally, side-road-based systems use sensors, such as magnetometers, accelerometers, and acoustic sensors, installed on a roadside . Recently, Distributed acoustic sensing (DAS) is emerging as tool for collecting space-time data without requiring a dedicated installation of sensing equipment. DAS transforms a conventional optical fiber cable into a dense array of strain seismometers with detection points spaced every few meters along the fiber, allowing to monitor tens of kilometres. Further details on the physics and technologies underlying DAS are available in [3] and references therein. DAS is having a large number of applications such as train tracking, landslide detection, seismic activity, earthquake monitoring, among others.

In this short paper we introduce the fundamentals of a strategy for detecting vehicle passage and typology based on the analysis of DAS data recording the strain and deformation of a real world optic fiber, due to the passage of vehicles. Some recent papers ([2, 4]) have introduced interesting approaches for the analysis of DAS data with the aim of monitoring traffic flow however, they focus mostly on counting the vehicles on the road and compute their speed. As opposed to these papers, the presented strategy aims at not just counting, but evaluating the type of the passing vehicles.

While a dedicated fiber installation with a very high signal/noise ratio and no ambient noise should be ideal for addressing our challenge, the ability to use pre-existent (and non-ideal) fiber installations would provide useful advances for the deployment of DAS. This motivates our focus on real world DAS data and our experiments.

We have considered data recorded through an optic fiber installed along the road as shown in figure 1. The raw data collected by the central unit is a space-time matrix in which each row represents the observations along a specific section of the fiber and columns record measurements over the time. The dataset used is just a sample of the whole experiment and it comprehends 1081 observations collected over a 6 minutes and 40 seconds period of time, with a sampling frequency of $78Hz$. The most challenging issue of the used data set is that being data collected in an unconstrained environment, records include a lot of environmental noise. Moreover, the high volume of the acquired data require the use of appropriate dimensionality reduction techniques to support data transmission on reduced bandwidth networks and almost real-time processing.

The strategy we propose is based on processing data batches through three steps:

- 2D wavelet processing for noise filtering;
- Peaks detection for identifying the vehicle passage;
- waveform clustering for detecting the vehicle typology.

The next section summarizes the main steps and provide some results on data.

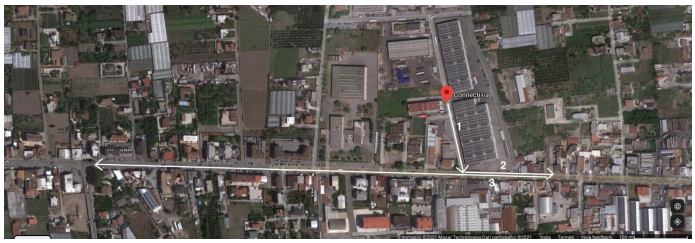


Fig. 1 Fiber positioning. The red spot indicates where the fiber starts.

2 Distributed acoustic sensing data processing

To analyse the data recorded by the distributed acoustic sensor, we monitor the optic fiber at n observation points $i = 1,\dots,n$. For each i, we have a series $Y_i = \{y_i^1,\dots,y_i^t,\dots,y_i^T\}$ which records the optic fiber strain due to vehicle passage in the time frame $t = 1,\dots T$.

We organise the DAS data in a matrix $Y_{n\times T} = \{Y_1,\dots,Y_i,\dots,Y_n\}$.

We assume the optical fiber to be placed parallel to the road so that the noise emitted by a moving vehicle is sensed over the time by consecutive sections of the fiber. That is, if the series Y_i records some information about the passage of a vehicle, the series Y_{i+1} also records it with some phase shift depending on the vehicle speed. Of course, as soon as the vehicle leaves the monitored road, no series will record its movement.

The first step of our strategy consists in performing a 2-dimensional wavelet decomposition of the matrix Y. Wavelet analysis allows to analyse signals and images at different resolutions, to detect change points, discontinuities, and other events not readily visible in raw data. A key advantage it has over other signal processing techniques, e.g. Fourier transform, is temporal resolution: it captures both frequency and location information (location in time).

The 2-dimensional wavelet transform allows to get a multi-level decomposition of the matrix Y such that for each level we have an approximation of the matrix and three sets of coefficients: horizontal, vertical, and diagonal coefficients. It is interesting to note that each set of coefficients highlights specific features of the data matrix.

In our specific application, since the trace of vehicle passage is recorded by phase shifted signals, we are interested in using the information captured by the diagonal coefficients. These are represented as a matrix $Z_{n',T'}$ whose dimension depends on the levels selected for the wavelet transform.

In figure 2 we show the results of processing data through the 2D wavelet transform. The left side represents a snapshot of the data as they are sensed by the DAS system. A visual inspection does not allow to detect any pattern, as expected. The right side of the figure represents the diagonal coefficient of the 2D wavelet transform. Here we see the passage of heavy vehicles as diagonal traces, confirming the usefulness of this kind of pre-processing.

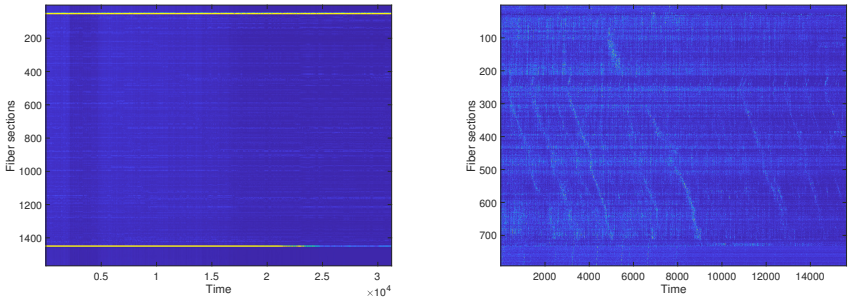


Fig. 2 Raw data vs diagonal coefficient of 2D wavelet transform

The second step we perform is the detection of peaks on the coefficient matrix. To this aim we use the classic short-term average/long-term average (STA/LTA) algorithm [1].

Our idea is to use the time stamp of each peak as the center of a time window so that each signal Z_i is represented by a set of sub-sequences detected through the selection of data around the peak.

Formally, for each Z_i, we detect the peaks $PK_i = \left\{pk^j\right\}_{j=1,\ldots,l_i}$, where pk^j is the time stamp of a peak. By considering a time window having size w, we can recover for each pk^j a time interval $\left[a^j = pk^j - w/2; b^j = pk^j + w/2\right]$ and the corresponding subsequence $s_i^j = \left\{z_i^{a^j}, \ldots, z_i^{b^j}\right\}$.

The clustering of sub-sequences allows to allocate vehicles to typologies. We use an algorithm based on BIRCH [6] since it is very effective in analysing huge amounts of data. A first phase of the algorithm obtains a partition of data into a high number of low variability clusters, performing a single scan. A second phase consists in running a k-means algorithm on the centroids of the clusters discovered by the first phase. The final reduced set of clusters corresponds to vehicle typologies.

The pseudo-code of the first phase of the algorithm is the following:

Initialization:
$i = 1$
Detect the peaks PK_i
Detect the sub-sequences s_i^j for each peak pk^j
Run a k-means algorithm on the s_i^j $(j = 1, \ldots, l_i)$ to get a partition in K clusters C_k and the centroids G_k
Main:
for all Z_i such that $i > 1$ **do**
    Detect the peaks PK_i
    Detect the sub-sequences s_i^j for each peak pk^j
    for all s_i^j **do**
        Allocate s_i^j to the cluster C_k such that:

$d^2(s_i^j; G_k) < d^2(s_i^j; G'_k)$ if $d^2(s_i^j; G_k) < u$

end for

Update the cluster centroids G_k $(k = 1, \ldots, K)$

end for

We have tested this algorithm on our dataset. In Fig. 3 we show the strain of the optic fiber at the observation point $i = 400$. The peaks represent the passage of vehicles. The coloured dots show the membership of each vehicle to a typology as result of a clustering of data into 4 clusters.

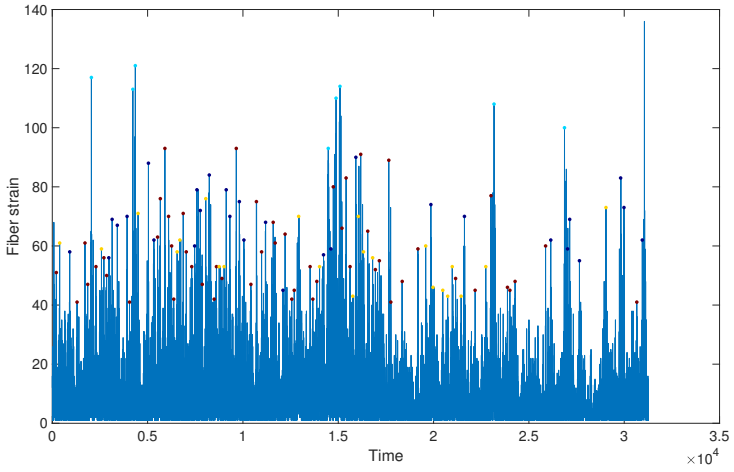


Fig. 3 Plot of the optic fiber strain for $i = 400$. The coloured dots represent the vehicle type provided as output by the clustering procedure

3 Conclusions

In this short paper we have shown a strategy for detecting the passage of vehicles and for clustering such vehicles into typologies starting from DAS data. We have run our algorithms on real data to evaluate the effectiveness of the strategy. By means of a camera, we have recorded the vehicle passage on the road to compare our results with the ground truth. Despite the data being affected by noise, we were able to obtain encouraging results. Further validations and tests to validate the input parameters of the procedure will be the subject of future works.

References

1. Allen, R. V.: Automatic earthquake recognition and timing from single traces. Bulletin of the Seismological Society of America, **68**, 1521–1532 (1978).
2. Chambers, K.: Using DAS to investigate traffic patterns at Brady Hot Springs, Nevada, USA. The Leading Edge, **39**, 819–827, (2020).
3. Hartog, A. H.: An Introduction to Distributed Optical Fibre Sensors. CRC Press (2017).
4. Liu, H., Ma, J., Yan, W., Liu, W., Zhang, X., Li, C.: Traffic Flow Detection Using Distributed Fiber Optic Acoustic Sensing. IEEE Access, **6**, 68968–68980, (2018).
5. Won, M.: Intelligent Traffic Monitoring Systems for Vehicle Classification: A Survey. IEEE Access, **8**, 73340–73358, (2020), doi: 10.1109/ACCESS.2020.2987634.
6. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96. pp. 103–114. (1996) doi:10.1145/233269.233324.

Session of solicited contributes SS18 – *Evaluation and assessment of cognitive and learning processes*
Organizer and Chair: Francesco Palumbo

Neuropsychological Assessment supported by Technology: the E-BTT case

Valutazione in ambito neuropsicologico supportato dalla tecnologia: l'esempio di E-BTT

Michela Ponticorvo and Antonietta Argiuolo

Abstract The assessment in neuropsychology is moving from the traditional paper-and-pencil tests to the technological enhanced ones. Introducing technology allows to open the way to new analyses thus expanding the assessment beyond diagnosis to explore spatial exploration. Here we describe the E-BTT, the technological enhanced version of BTT, a well-known and widely used test to assess neglect and show how this version that joins digital and physical elements exploits technology potential in assessment.

Abstract *La valutazione in neuropsicologia si sta spostando dai tradizionali test carta e matita a quelli tecnologicamente avanzati. L'introduzione della tecnologia consente di aprire la strada a nuove analisi ampliando così la valutazione oltre la diagnosi per esplorare l'esplorazione spaziale. Qui descriviamo l'E-BTT, la versione tecnologicamente avanzata di BTT, un test noto ed ampiamente utilizzato per valutare il neglect e mostrare come questa versione che unisce elementi digitali e fisici sfrutti il potenziale della tecnologia nella valutazione.*

Key words: neuropsychology assessment, spatial exploration, technology-enhanced assessment

1 Neuropsychological Assessment of Spatial Disorders

Assessment in neuropsychological is crucial as it allows to diagnose disorders related to injuries, strokes, decay etc. It is devoted to different cognitive areas, including memory (Loring & Papanicolaou, 1987), executive functions (McCloskey & Perkins, 2012), spatial abilities etc.

Regarding spatial abilities there are various tools, validated and widely used in clinical practice that allow to discriminate between preserved and compromised spatial abilities (Cerrato, Ponticorvo, Gigliotta, Bartolomeo and Miglino, 2019a; Azouvi, Bartolomeo, Beis, Perennou, Pradat-Diehl and Rousseaux, 2006; Doricchi and Bartolomeo, 2018). In general, these are paper- and-pencil tasks that require the subject or patient to explore, process and locate stimuli: this is the case, for example, of cancellation tests (of lines, letters or other symbols), where it is necessary to cross out some target stimuli with or without distractors (e.g. the bell test, line barrage); or bisection of lines, in which segments of different lengths are presented and the task is to find and mark the midpoint. Other widely used tests are the copying of drawings (such as the complex figure of Rey-Osterrieth; Rey, 1941; Osterrieth, 1944), the clock test (Agrell and Dehlin, 1998) and the free drawing.

1.1 Some limitation of traditional tools

The traditional tools cited above have some limitations, especially related to sensitivity. As discussed in Cerrato and colleagues (2020), it is often noted that scores within the norm, especially in chronic patients with neglect in apparent remission, are not always accompanied by a real functional recovery of the patient, but rather by the implementation of a series of compensation strategies. Takamura and colleagues (2016) describe how patients who recognize their neglect intentionally focus on the neglected space as a compensatory strategy. The massive presence of these compensatory strategies makes it difficult to discriminate between a real recovery and the implementation of such strategies (Bonato, 2012).

Another weak point of the traditional paper-and-pencil tests is their scarce ecological validity, that is, their little generalizability to situations other than those of the clinic (Lewkowicz, 2001); to improve it, instead, it is necessary to propose actions performed in daily life, such as combing one's hair or powdering the Comb and Razor Task (Beschin and Robertson, 1997); or how to arrange objects in space, as is done when cookies need to be baked (Cerrato et al., 2019a b; Tham and Tegnér, 1996): this is what is required for the test we are introducing in the next section.

2 The Baking Tray Test and its enhanced version E-BTT

The Baking Tray Task, also known with its acronym BTT, is a versatile and ecological neuropsychological test originally developed by Tham and Tegner (1996). It arises from the need for an alternative to the classic paper and pencil tests for the clinical evaluation of neglect, as these are widely used and validated, but show strong ecological and sensitivity limits, as discussed above.

It is more valid at ecological level to ask to participants to objects in a peripersonal space, as is done when cookies need to be baked.

The latter activity inspired the BTT. As the name suggests, it simulates a daily situation with a board of 75x100 cm (the “tray”) and 16 small cubes of 3.5 cm (the “sandwiches” or “biscuits”), asking the participant to arrange them “as evenly as possible on the whole table, as if they were sandwiches on a tray to be baked” (Tham and Tegnér, 1996; p. 20). An unbalanced spatial arrangement between right and left of more than 2 cubes was considered by the authors to be pathological (and therefore an index of hemi-neglect) see Facchin and colleagues (2016).

The BTT offers significant advantages, it is easy to use and administer; it is relatively fast and requires a relatively light attention load; despite this, it appears to be able to correctly identify patients with neglect, while other tests can lead to false negatives (Halligan and Marshall, 1992). Thanks to the unspecified nature of the spatial arrangements of the cubes, it would also seem to be free from effects due to practice, which instead affect performance in cancellation tests (Tham and Tegnér, 1996), or from demographic variables such as age, education and gender (Facchin, Beschin, Pisano and Reverberi, 2016). The BTT also showed good test-retest reliability (Bailey, Riddoch and Crome, 2004).

Recently, Cerrato and colleagues (2019a; 2019b) developed an improved and enhanced version of the BTT based on E-TAN, a platform that supports tangible interfaces (Cerrato et al., 2020; Gentile, Cerrato, Ponticorvo, 2019) to investigate visuospatial behaviors in the proximal or peripersonal space. This is the tool we are presenting here, the E-BTT (previously called BTT-SCAN), which replaces the cubes with 4 cm diameter discs and the "tray" with a surface of 60 x 45 cm bordered by a wooden frame.

On the disks there are tags (ArUco Markers by Garrido-Jurado, Muñoz-Salinas, Madrid-Cuevas & Marín-Jiménez, 2014) containing a QR code. A computer vision software receives the tag signal through the video camera, placed perpendicular to the frame, which scans and locates the diskettes, recording all their movements. This

mechanism makes it possible to accurately collect the coordinates of the discs and to analyze the strategies for their arrangement within the frame.

The E-TAN platform falls into the category of tangible user interface systems (Ferrara, Ponticorvo, Di Ferdinando and Miglino, 2016), consisting of an integrated system of concrete objects that participants can manipulate, thus maintaining the ecological value of BTT's traditional task and at the same time enhancing it by increasing its information capacity thanks to the union of the digital system.

3 Assessment supported by technology beyond diagnosis

The tool we have described E-BTT, represented in figure 1, offers the great advantage to record much richer information that traditional BTT First of all, it is possible to collect Cartesian coordinates of each disk, as well as their temporal sequence (Cerrato et al., 2019b; Gentile et al., 2019; Palumbo, Cerrato, Ponticorvo, Gigliotta, Bartolomeo and Miglino, 2019a). Such an innovation allows not only to discriminate between pathological and normal disposition, but also to investigate the exploration trajectories in healthy subjects, as well as the area occupied in the arrangement of objects in space.

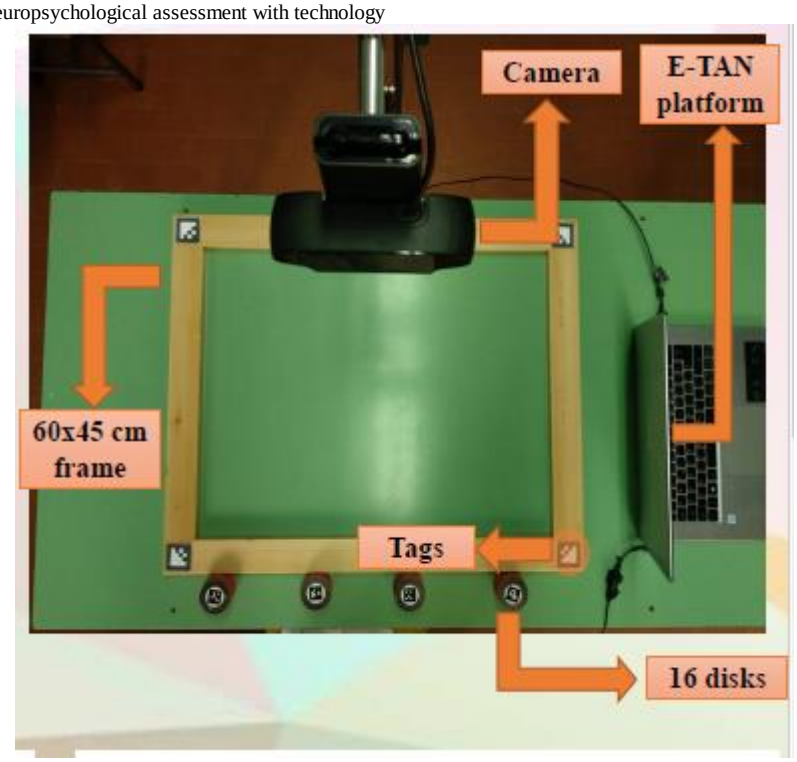


Figure 1: E-BTT with digital and tangible interfaces

The tests presented before, indeed, arise in the clinical setting, to detect the presence and trend of spatial cognition deficits. However, they have also been widely used with healthy people, highlighting the presence of pseudoneglect (Gigliotta et al., 2017; Cerrato et al., 2019a), i.e. the tendency to deviate to the left. More in general it can be applied to detect tendency to explore the space, considering laterality, as well as verticality and organization of spatial exploration (Argiuolo & Ponticorvo, 2020).

References

1. Agrell, B., Dehlin, O.: The clock-drawing test. Age and ageing, 27(3), 399-403 (1998).
2. Albert, M.: A simple test of visual neglect. Neurology, 23, pp. 658-664 (1973).
3. Azouvi, P., Bartolomeo, P., Beis, J.-M., Perennou, D., Pradat-Diehl, P., Rousseaux, M.: (2006). A battery of tests for the quantitative assessment of unilateral neglect. Restorative neurology and neuroscience, 24(4-6), pp. 273–285 (2006).

4. Bailey, M.J., Riddoch, M.J., Crome, P.: Test–retest stability of three tests for unilateral visual neglect in patients with stroke: Star cancellation, line bisection, and the baking tray task. Neuropsychological Rehabilitation, 14(4), pp. 403–419 (2004).
5. Beschin, N., Robertson, I.H.: Personal Versus Extrapersonal Neglect: A Group Study of their Dissociation Using a Reliable Clinical Test. Cortex, 33(2), pp. 379–384 (1997).
6. Bonato, M.: Neglect and extinction depend greatly on task demands: a review. Frontiers in human neuroscience, 6, 195 (2012).
7. Cerrato, A., Pacella, D., Palumbo, F., Beauvais, D., Ponticorvo, M., Miglino, O., et al.: E-TAN, a technology-enhanced platform with tangible objects for the assessment of visual neglect: A multiple single-case study. Neuropsychological Rehabilitation, pp. 1-15 (2020).
8. Cerrato, A., Ponticorvo, M., Gigliotta, O., Bartolomeo, P., Miglino, O.: Btt-scan: uno strumento per la valutazione della negligenza spaziale unilaterale. Sistemi intelligenti, 31(2), pp. 253–270 (2019a)
9. Cerrato, A., Ponticorvo, M., Gigliotta, O., Bartolomeo, P., Miglino, O.:. The assessment of visuospatial abilities with tangible interfaces and machine learning. In International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer: Berlino, Heidelberg, pp. 78–87 (2019b).
10. Doricchi, F., Bartolomeo, P.: Neuropsicologia dell'attenzione. In G. Denes, L. Pizzamiglio, C. Guariglia, S. Cappa, D. Grossi, e C. Luzzatti (a cura di), Manuale di Neuropsicologia. Bologna: Zanichelli, pp. 705-728 (2018).
11. Facchin, A., Beschin, N., Pisano, A., Reverberi, C.: Normative data for distal line bisection and baking tray task. Neurological Sciences, 37(9), pp. 1531–1536 (2016).
12. Ferrara, F., Ponticorvo, M., Di Ferdinando, A., Miglino, O.: Tangible interfaces for cognitive assessment and training in children: Logicart. In V. L. Uskov, R. J. Howlett, e L. C. Jain (a cura di), Smart education and e-learning. Berlino, Heidelberg: Springer, pp. 329–338 (2016).
13. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., Marín-Jiménez, M. J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition, 47(6), 2280–2292 (2014).
14. Gentile, C., Cerrato, A., Ponticorvo, M.: Using technology and tangible interfaces in a visuospatial cognition task: the case of the baking tray task. In Miglino, O. & Ponticorvo, M. (a cura di), Proceedings of the First Symposium on Psychology-Based Technologies. CEUR Workshop Proceedings (2019).
15. Gigliotta, O., Malkinson, T.S., Miglino, O., Bartolomeo, P.: Pseudoneglect in visual search: Behavioral evidence and connectional constraints in simulated neural circuitry. eNeuro (2017).
16. Halligan, P.W., Marshall, J.C.:. Left visuo-spatial neglect: A meaningless entity?. Cortex, 28(4), pp. 525–535 (1992).
17. Lewkowicz, D.J.: The Concept of Ecological Validity: What Are Its Limitations and Is It Bad to Be Invalid?. Infancy, 2, pp. 437-450 (2001).
18. Loring, D. W., & Papanicolaou, A. C.: Memory assessment in neuropsychology: Theoretical considerations and practical utility. Journal of Clinical and Experimental Neuropsychology, 9(4), 340-358 (1987).
19. McCloskey, G., & Perkins, L. A.:(2012). Essentials of executive functions assessment (Vol. 68). John Wiley & Sons.
20. Osterrieth, P. A.: Le test de copie d'une figure complexe. Archives de Psychologie, 30, 205-550 (1994).
21. Palumbo, F., Cerrato, A., Ponticorvo, M., Gigliotta, O., Bartolomeo, P., Miglino, O.: Clustering of behavioral spatial trajectories in neuropsychological assessment. In SIS 2019-Smart Statistics for Smart Applications. Pearson: Londra, UK, pp. 463–470 (2019).
22. Rey, A.: L'examen psychologique dans les cas d'encéphalopathie traumatique. Archives de psychologie (1941).
23. Argiuolo, A., & Ponticorvo, M.: E-TAN platform and E-baking tray task potentialities: new ways to solve old problems. In PSYCHOBIT, Ceur Proceedings, 2730 (2020).
24. Takamura, Y., Imanishi, M., Osaka, M., Ohmatsu, S., Tominaga, T., Yamanaka, K., et al.: Intentional gaze shift to neglected space: a compensatory strategy during recovery after unilateral spatial neglect. Brain, 139(11), pp. 2970-2982 (2016).

Tham, K. Tegnér, R.: The baking tray task: a test of spatial neglect. Neuropsychological Rehabilitation, 6(1), pp. 19–26 (1996).

Introducing OpenAI-ES in Interactive Data Clustering with R-EVOK

Clusterizzazione interattiva in R-EVOK attraverso OpenAI-ES

Nicola Milano and Onofrio Gigliotta

Abstract Data clustering allows to identify homogenous group within a dataset. Data clustering techniques are widely used in the context of machine learning and often are the engine under the hoods of many applications. Bio-inspired algorithms and classical Evolutionary Computation techniques, although able to reach comparable solutions to other mainstream algorithms, make use of bigger computational power. Recently, the interest on Evolutionary Algorithms has produced new solutions able to compete with state- of- the- art algorithms in tasks requiring machine learning. In this paper we describe how OpenAI-ES, a novel evolutionary algorithm, can be used to boost human performance in R-EVOK an interactive clustering software in which a human being (acting as a breeder) selects genetically encoded cluster configurations graphically represented by Rousseeuw's Silhouttes (phenotypes). In this work we present a first implementation of the algorithm and report preliminary results.

Abstract *Il clustering dei dati consente di identificare un gruppo omogeneo all'interno di un set di dati. Le tecniche di clustering dei dati sono ampiamente utilizzate nel contesto dell'apprendimento automatico e spesso sono il motore di molte applicazioni. Algoritmi bio-ispirati e tecniche classiche di calcolo evolutivo, sebbene in grado di raggiungere soluzioni comparabili ad altri algoritmi tradizionali, fanno uso di una maggiore potenza computazionale. Recentemente, l'interesse per gli Algoritmi Evolutivi ha prodotto nuove soluzioni in grado di competere con algoritmi all'avanguardia in compiti che richiedono l'apprendimento automatico. In questo articolo descriviamo come OpenAI-ES, un nuovo algoritmo evolutivo, può essere utilizzato per aumentare le prestazioni in R-EVOK, un software di clustering interattivo in cui un essere umano (che agisce come un selezionatore) seleziona configurazioni di cluster codificate geneticamente rappresentate graficamente da Rousseeuw Silhoutte (fenotipi). In questo lavoro*

presentiamo una prima implementazione dell'algoritmo e riportiamo i risultati preliminari

Key words: clustering, evolutionary algorithms, Rousseeuw Silhoutte

1 Introducing Evolutionary Computation

Evolutionary Computation, in data clustering, has been supported by a solid scientific literature [1]. When using EC we must define a fitness function that, in the case of data clustering, have to measure the quality of the clusters. This function can be formalized as mathematical formula or even delegated to a human user as in REVOK a bio-inspired interactive evolutionary algorithm developed by Russo and colleagues [2]. Cluster analysis plays a crucial role in many knowledge fields. Biological learning/adapting processes naturally produce clusters of objects, perceptions, concepts etc. Since clustering represents an important base in order to make decisions in different scientific contexts, a variety of methods have been developed [3]. Evolutionary algorithms are effective metaheuristics able to tackle NP-hard problems and clustering can be considered one of those [4]. Moreover, this approach allows us to completely automate the search of the most suitable number of clusters within a specific dataset.

2 R-EVOK

R-EVOK [2] one of the present authors contributed by designing the underlying interactive genetic algorithm (GA). In that work each cluster, encoded genetically, could be silenced or activated by a single mutation. The interactive fitness function was implemented by showing users the silhouette of 9 different data clustering. Users then were asked to select 3 of them to undergo the evolutionary process, a process consisting in mutating and replicating each individual solution. This approach has the merit to exploit the knowledge of the user but can be difficult to manage. For this reason, in this paper, we present a completely automated system able to find, by means of interactive evolutionary computation, the most suitable number of clusters in a dataset.

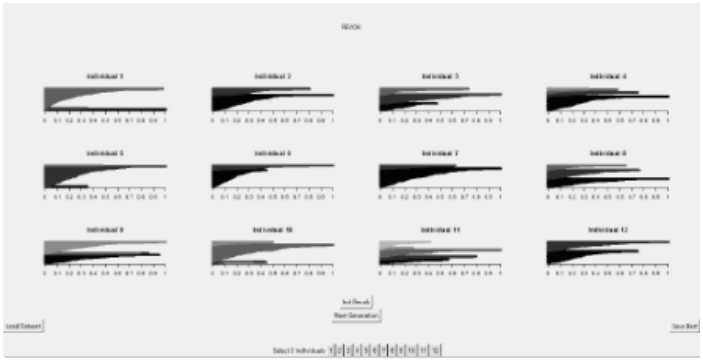


Figure 1:R-EVOK wondow displaying Silhouttes for each clustering process

3 OpenAI ES and clustering

OpenAI-ES method is one of the most effective evolutionary technique [4]. It consists in creating a population starting from a single parent (theta). Then, the current fitness (after being ranked to avoid the effect of extreme values) of each generated individual is used to estimate the gradient of the expected fitness. The latter is then used to recompute the centre of the distribution, i.e. the parent theta coordinates by using the Adam stochastic optimizer. This process is reiterated for a number of generations that can be set by the user.

Every individual is encoded in a genetic string representing the centroid of each possible cluster (the maximum can be defined by the users, see figure 2). Centroids can be active or inactive depending on the value of a controlling gene. The entire genetic string, denoting a set of possible clusters, undergoes to the evolutionary process.

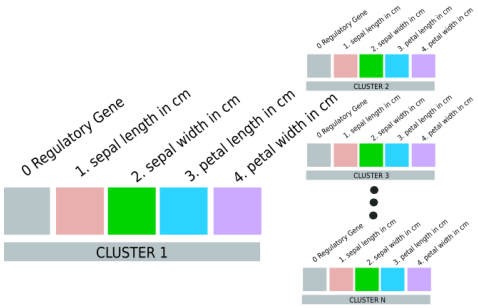


Figure 2 : Genetic string in the case of Iris dataset

4 Preliminary results and Conclusion

R-EVOK and R-EVOK with OpenAI-ES have been tested with two datasets: Iris and Haberman downloaded from the UCI machine learning repository (http://archive.ics.uci.edu/ml/). We recruited 15 undergraduate students and asked them to partition the two datasets according only to aesthetic/qualitative criterion (after a brief explanation how Silhouettes represent good partitions). Individuals interacting with the novel OpenAI-ES powered version of R-EVOK tend to find good solutions in fewer interactions. The advantage of such an approach, in general, relies on the possibility to ground a data clustering process onto the knowledge of the user that very often could be unconscious [9] or simply implicit. Uniting human competences and machine learning power, to us, is the future direction to follow. Hence future versions of the present software should take in account how to better extract implicit/unconscious knowledge of users to improve the grounding value of a specific clustering.

References

1. Abul Hasan, M.J., Ramakrishnan, S.: A survey: hybrid evolutionary algorithms for cluster analysis. Artificial Intelligence Review **36** (3), 179–204 (2011)
2. Russo, A., Gigliotta, O., Palumbo, F., Miglino, O.: Introducing Interactive Evolutionary Computation in Data Clustering
3. Everitt, B., Landau, S., Leese, M.: Cluster Analysis. A Hodder Arnold Publication, Wiley (2001)
4. Milano, N., Nolfi, S. Automated curriculum learning for embodied agents a neuroevolutionary approach. Sci Rep **11,** 8985 (2021). https://doi.org/10.1038/s41598-021-88464-5
5. Hruschka, E. R., Campello, A. A. Freitas and A. C. Ponce Leon F. de Carvalho, "A Survey of Evolutionary Algorithms for Clustering," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 39, no. 2, pp. 133-155, March 2009, doi: 10.1109/TSMCC.2008.2007252.
6. Kihlstrom, J.: The cognitive unconscious. Science **237**(4821), 1445–1452 (1987)

Modeling heterogeneity in students' satisfaction during the Covid-19 pandemia

Analisi delle fonti di eterogeneità nella soddisfazione degli studenti durante la pandemia Covid-19

Cristina Davino, Marco Gherghi, Francesco Palumbo and Domenico Vistocco

Abstract Aim of the paper is to analyse whether and to what extent certain drivers of student satisfaction on teaching activities have different effects depending on the satisfaction levels and characteristics of the students. The analysis focuses on the effects of distance learning suddenly introduced in Italian universities because of the Covid-19 pandemic. An analysis strategy based on the use of quantile regression is proposed as a solution for identifying different models in the case of observed heterogeneity.

Abstract *L'obiettivo del lavoro è analizzare se e quanto le motivazioni che regolano la soddisfazione degli studenti rispetto alla didattica hanno effetti diversi in base ai differenti livelli di soddisfazione e alle differenti caratteristiche degli studenti. L'analisi si concentra sugli effetti della didattica a distanza introdotta nelle università italiane a seguito della pandemia da Covid-19. In particolare si propone una strategia di analisi basata sull'uso della regressione quantile come soluzione per l'individuazione di modelli diversi in caso di presenza di eterogeneità.*

Key words: quantile regression, e-learning, student satisfaction

Cristina Davino
Department of Economics and Statistics, University of Naples Federico II e-mail: cristina.davino@unina.it

Marco Gherghi
Department of Economics and Statistics, University of Naples Federico II e-mail: gherghi@unina.it

Francesco Palumbo
Department of Political Science, University of Naples Federico II e-mail: fpalumbo@unina.it

Domenico Vistocco
Department of Political Science, University of Naples Federico II e-mail: domenico.vistocco@unina.it

1 Introduction and reference framework

The Covid-19 pandemic disrupted lives around the world and imposed new ways of carrying out activities crucial to people's lives and the progress of society. It is well known that education is one of the sectors most affected by the restrictions on individual freedom needed to contain the virus. Indeed, in order to avoid the complete suspension of training activities, it was necessary to adopt a teaching method totally different than the one traditionally used, namely distance learning. While the enormous effort made to quickly switch to this new teaching method has been appreciated by students and their families, the transition has certainly had a major impact on the lives of involved teachers, students and families. This impact was inevitably conditioned, positively or negatively, by the personal characteristics of the students and the socio-economic conditions of the families, with the obvious consequence of increasing inequalities [2]. An open issue concerns the evaluation of the effectiveness of e-learning and the need/opportunity to use it still in the future [4] [8]. Such an evaluation must consider different elements such as the family impact of the e-learning experience, the digital divide, the effects of different forms of e-learning (e.g. synchronous and asynchronous), the capacity of educational institutions to organise and adapt to the emergency and, last but not least, students' evaluations and perceptions of the experience.

This paper focuses on the higher education sector and aims to investigate the determinants of students' satisfaction about the e-learning experience by exploring whether the drivers of satisfaction change according to student characteristics. The proposed study is based on the idea that the possible drivers of satisfaction can affect differently poor, moderate or very satisfied students. The use of quantile regression (QR) [6, 1, 5] as a complementary tool to classical ordinary least squares regression (OLS) allows to capture the effects of the set of considered regressors on the entire conditional distribution of the dependent variable. Moreover, one or more variables not considered in the regression model could affect the strong, and sometimes the verse, of the dependence structure. Such variables, also called stratification variables, describe the membership of the units to different groups (the different levels of the variables) and are a source of possible heterogeneity in the estimated model. Among the different approaches proposed in the literature to analyze group effects in a dependence model it is possible to list the use of separate models for each group, the use of dummy variables into the model, and the more refined approach based on multilevel models. Here, we exploit the potentialities of the approach proposed by Davino and Vistocco [3] to identify group effects through a quantile regression model. The method assigns a conditional quantile to each group and provides a separate analysis of the dependence structure inside the groups. The approach is structured in three steps: identification of the best model for each group, estimation of the group dependence structure and test of the differences among groups. In the first step, the conditional quantiles representative of each group are determined by computing the rank percentiles of each statistical unit with respect to the response variable and then averaging them by groups. In the second step, QR is carried out on the whole sample using the quantiles assigned to the groups in the previous step.

The obtained QR coefficients highlight possible differences among the groups. In the final step, the evaluation of the statistical significance of the differences among the coefficients related to each group is performed through classical inferential tools available in the quantile regression framework [6, 1, 5]. It is important to highlight that group coefficients can be compared because they have been estimated on the whole sample, unlike approaches estimating separate models for each group. This final step can be carried out using one of the classical tests proposed in [7] and aimed at evaluating the significance of the differences among coefficients pertaining to different quantiles. The most common test statistic is a variant of the Wald test, which provides a joint test on all slope parameters. In case of refusal of the null hypothesis of no difference among groups, pairwise comparisons is used to investigate the difference structure between all the possible pairs of groups.

2 Data description and main results

The study proposed in this paper is based on a survey administered at the University of Naples and considering students (10,239 participants) who attended at least one distance learning course in the 2019/2020 academic year. The observed sample reflects the distribution of the student population by degree course.

In this paper we will focus on the evaluations provided by students on elements that are known to have an impact on satisfaction (*score*): the family impact (*FIEL*) of the e-learning experience (i.e. organisation of space at home, expenses incurred), the organisation (*organisation*) of all complementary activities (exams, reception, dissertation), the technological equipment (*tech_equipment*) and any technical problems (*tech_problems*) encountered during online lessons. These drivers of satisfaction, as well as the satisfaction itself, are composite indicators obtained through a multiple correspondence analysis of the questions relating to each topic. To facilitate the interpretation of the results, the variables are scaled in the range 0-100.

Aim of the empirical analysis is to model the variable *score* through the regressors *FIEL*, *organisation*, *tech_equipment* and *tech_problems* following a twofold objective: to explore whether these regressors have different effects on different levels of satisfaction, and study whether this heterogeneity is also related to personal characteristics of students (gender, age, degree course and residence, in particular). Due to lack of space, we will only present the results differentiated by age groups. The observed sample consists of the following groups: age 18–21 (57%), age 22–23 (23%), and age higher than 23 years (20%). Figure 1 shows the overall distribution of the variable *score* and its distribution by age groups. It is noteworthy to highlight the asymmetrical form of this variable and the increasing levels of satisfaction as age increases, at least for the 50% of students in the central part of the distribution.

The quantiles (θ_{best}) representative of each age group are: 0.45, 0.52 and 0.63, respectively. The identified best quantiles characterize the groups, meaning that, for example, the dependence structure for younger students (18–21 years) is best rep-

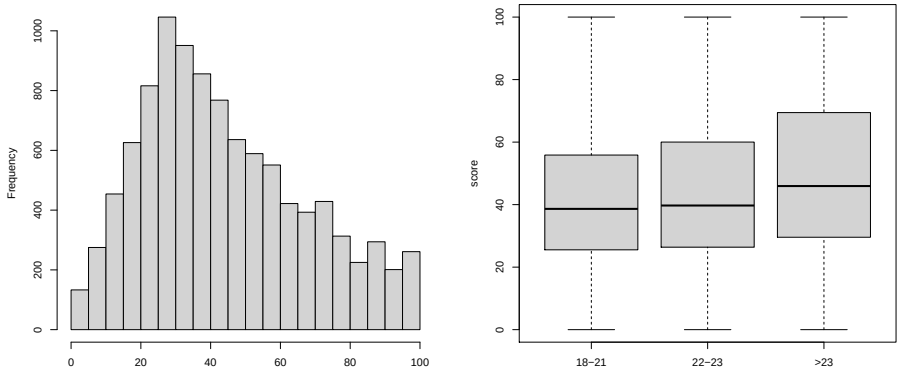


Fig. 1 Distribution of the *score* on the whole sample (left-hand side) and for age groups (right-hand side)

resented by a QR model estimated at the quantile 0.45, i.e. the features of younger students mainly affect the 45th conditional percentile of the *score*.

The comparison among the best quantiles values heralds possible group differences and peculiarities. In the second step, QR is carried out on the whole sample using the best quantile assigned to the groups in the previous step. Table 1 reports the results of the three quantile regression models for each θ_{best} (second to fourth column) and the results of the classical OLS model (last column). There is an upward trend, as age increases, in the effect (negative) of family impact, and in the effect (positive) of assessment on the organisation of all activities complementary to lessons (exams, thesis, student reception). The effect of technological equipment is low, although still significant.

Table 1 Group effects and OLS (last column) results (significant coefficients in bold, $\alpha = 0.05$). The coefficients measure the dependence structure inside each group.

	θ_{best}=0.45	θ_{best}=0.52	θ_{best}= 0.63	OLS
(Intercept)	**24.987**	**27.465**	**33.376**	**31.701**
FIEL	**-0.143**	**-0.151**	**-0.165**	**-0.155**
organization	**0.545**	**0.595**	**0.614**	**0.483**
tech_equipment	**0.025**	**0.024**	**0.030**	**0.034**
tech_problems	-0.006	-0.006	-0.008	-0.003

In the final step, the evaluation of the statistical significance of the differences among the coefficients related to the three groups is carried out, results shown in Table 2 . In particular, in case the difference among the coefficients of the three

groups is significant (second column of Table 2), pairwise comparisons between groups allow investigate the structure of the differences. Starting from the three groups of age above described, 3 possible comparisons are possible. Columns from second to fourh Table 2 report the *p*–values deriving from testing differences on each slope coefficient, for each pair of models. It is worth to recall that the null hypothesis states that the slope coefficients of the two models are identical. We refer to the three models estimated in the previous step with G1, G2 and G3 in Table 2. Significant differences come to light for the regressors *FIEL* and *organization*, both on the whole model and for couples of models.

Table 2 P-values derived from testing differences on each slope coefficient (rows) obtained considering all the possible pairwise comparisons between groups and on the whole model (last column)

	joint test	G1 vs G2	G1 vs G3	G2 vs G3
FIEL	0.055	0.134	0.016	0.069
organization	0.000	0.000	0.000	0.240
tech_equipment	0.776	–	–	–
tech_problems	0.977	–	–	–

The analysis presented in this short paper can be enriched by considering more complex models, i.e. by introducing other regressors that may impact on satisfaction but also by exploring the heterogeneity of relationships between different groups of students, for example with respect to gender, residence or type of course. These aspects will be considered in an extended version of the study.

References

1. Davino, C., Furno, M., Vistocco, D.: Quantile Regression: Theory and Applications. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Inc (2013).
2. Davino, C., Gherghi, M. Vistocco, D.: A quantitative study to measure the family impact of e-learning, In B. Bertaccini, L. Fabbris, A. Petrucci (eds) ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference, Firenze University Press, pp. 103-107 (2021)
3. Davino, C., Vistocco, D. Handling heterogeneity among units in quantile regression. Investigating the impact of students' features on University outcome. Statistics and ist Interface, vol. 11, p. 541-556 (2018)
4. Di Pietro, G., Biagi, F., Costa, P., Karpiski Z., Mazza, J.: The likely impact of COVID-19 on education: Reflections based on the existing literature and international datasets. EUR 30275 EN, Publications Office of the European Union, Luxembourg (2020).
5. Furno, M., Vistocco, D.: Quantile Regression: Estimation and Simulation. Wiley Series in Probability and Statistics. Wiley (2018).
6. Koenker, R.: Quantile Regression. Econometric Society Monographs No. 38. New York: Cambridge University Press. (2005).
7. Koenker, R., Basset, G.: Robust tests for heteroscedasticity based on regression quantiles. Econometrica **50**(1), 43–61 (1982).

8. Pokhrel, S., Chhetri, R.: A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. Higher Education for the Future **8**(1), 133–141 (2021)

Session of solicited contributes SS19 – *Substainability and Environment*
Organizer and Chair: Ida Camminatiello

Partitioning the Cressie-Read divergence statistic for three-way contingency tables: a study on environmental sustainability data.

Decomposizione delle statistiche divergenti di Cressie-Read per tabelle di contingenza a tre-vie. Test su dati di sostenibilità ambientale.

Rosaria Lombardo and Eric J. Beh

Abstract When studying the association between the variables of a three-way contingency table, Lancaster [10] proposed different partitions of Pearson's three-way chi-squared statistic. This statistic is a special case of the three-way generalisation of the Cressie-Read divergence statistic [6]. To test the association among environmental sustainability variables, this paper presents an additive orthogonal partition of the generalised Cressie-Read divergence statistic under the assumption of complete independence between the variables.

Abstract *Per l'analisi dell'associazione tra le variabili di una tabella di contingenza a tre-vie, Lancaster [10] propose differenti partizioni del chi-quadrato a tre-vie di Pearson. Questa statistica è anche vista come un caso particolare della statistica divergente di Cressie-Read [6], generalizzata per tabelle a tre-vie. Per verificare la significatività statistica dell'associazione tra alcune variabili della sostenibilità ambientale, questo lavoro presenta una partizione ortogonale additiva della statistica divergente di Cressie-Read generalizzata, sotto l'ipotesi di indipendenza completa tra le variabili.*

Key words: Orthogonal Partition, Cressie-Read divergence statistic, Testing association

1 Three-way Cressie-Read Divergence Statistic

To determine whether there exists a statistically significant association between the row, column and tube variables of a three-way contingency table, one may calculate

Rosaria Lombardo
University of Campania L. Vanvitelli, Gran Priorato di Malta, Capua (CE), e-mail: rosaria.lombardo@unicampania.it

Eric J. Beh
University of Newcastle, Australia e-mail: eric.beh@newcastle.edu.au

any number of measures. The most common that are used for such a purpose include the three-way generalisation of Pearson's chi-squared statistic [10, 5] and the log-likelihood ratio statistic. These statistics can be shown to be special cases of the three-way Cressie-Read divergence statistic [6, 14, 15, 16] as are generalisations of the Freeman-Tukey statistic [7], the modified chi-squared statistic [12, 13] and the modified log-likelihood ratio statistic [9]. All of these statistics, and other special cases of the divergence statistic, are chi-squared random variables with $(IJK-1)-(I-1)-(J-1)-(K-1)$ degrees of freedom.

Here, we propose a three-way extension of the Cressie-Read divergence statistic and then examine a method of orthogonally partitioning this statistic. Doing so provides a means of generalising to three categorical variables the benefits of the Cressie-Read divergence statistic. The partition can also be used to examine and test the association between three categorical variables under the assumption of complete independence and in the presence of sparse data. Further generalisations of the partition to the multi-way case can certainly be considered but we shall examine this extension at a later date.

Let $\underline{\mathbf{N}}$ be an $I \times J \times K$ three-way contingency table belonging to the space $\Re^{I\times J\times K}$, where the (i, j, k)th cell entry has a frequency of n_{ijk} for $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$. Define n the grand total of $\underline{\mathbf{N}}$ and let the matrix of relative frequencies be $\underline{\mathbf{P}}$ so that its (i, j, k)th cell entry is $p_{ijk} = n_{ijk}/n$ where $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = 1$. Define the ith row marginal proportion by $p_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}$. Similarly, let $p_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$ be the jth column marginal proportion, and $p_{\bullet\bullet k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$ the kth tube marginal proportion.

The Cressie-Read divergence statistic [6] has been extensively studied for two-way contingency tables and has been extended for studying the association in a three-way contingency table; see, for example, [15]. Such a divergence statistic is defined here as

$$\mathrm{CR}(\delta) = \frac{2n}{\delta(\delta+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}\left\{\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{\delta} - 1\right\}, \tag{1}$$

where $\delta \in (-\infty, \infty)$. The general nature of (1) ensures that specific values of δ lead to specific measures of association, all of which are chi-squared random variables.

The most common special cases of (1) were considered by [14]. In this paper, we focus our attention on some of those special cases. Specifically, we focus our attention on Pearson's chi-squared statistic (when $\delta = 1$), on the Cressie-Read statistic (when $\delta = 2/3$) and on the Freeman-Tukey statistic ($\delta = -1/2$) which are, respectively,

$$\mathrm{CR}(\delta = 1) = X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(p_{ijk} - p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\right)^2}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} \tag{2}$$

$$\mathrm{CR}(\delta = 2/3) = CR = \frac{9n}{5}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}\left[\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{2/3} - 1\right] \tag{3}$$

$$\mathrm{CR}\left(\delta=-\frac{1}{2}\right)=T^2=4n\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\sqrt{p_{ijk}}-\sqrt{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2. \tag{4}$$

Other measures, which are notable members of the family of Cressie-Read divergence statistics, can be generalised to three-way data and include the modified chi-squared statistic $N^2=CR(\delta=-2)$, the log-likelihood ratio statistic $G^2=CR(\delta=0)$ and its modified version $M^2=CR(\delta=-1)$.

In the context of goodness-of-fit testing for two categorical variables, Cressie and Read [6] examine the appropriate values of δ that one should use. For a two-way contingency table, such that $\delta\neq-1,0$, they also recommend that $\delta\in(0,3/2]$ when $n>10$ and $\min(np_{i\bullet}p_{\bullet j})>1$ for all $i=1,2,\ldots,I$ and $j=1,2,\ldots,J$. Cressie and Read [6] advised that an appropriate choice of δ is $2/3$ leading to their statistic; $\mathrm{CR}(\delta=2/3)=CR$. The selection criteria for δ for a three-way contingency table can be made in an analogous manner.

The term $p_{ijk}/(p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k})$ in (1) is defined as the *Pearson ratio* of the (i,j,k)th cell of the contingency table and is just the ratio of the observed cell frequency to what is expected under complete independence; see [2, p. 123] and [3, 8] for a definition of this ratio in the context of correspondence analysis. When this ratio is equal to 1 for all cells, (1), and hence its special cases (including X^2, G^2, T^2, N^2 and M^2), is zero providing evidence that the three variables of $\underline{\mathbf{N}}$ are completely independent. One advantage of considering the Pearson ratio's is that they ensure that the log-transformation of the cell's proportion, p_{ijk}, is "triple-centred" by the log-transformed row, column and tube proportions, i.e. $\ln\left(p_{ijk}/(p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k})\right)=\ln\left(p_{ijk}\right)-\ln(p_{i\bullet\bullet})-\ln(p_{\bullet j\bullet})-\ln(p_{\bullet\bullet k})$.

1.1 Partitioning the Cressie-Read divergence statistics

Here, we present an additive orthogonal, and ANOVA-like, partition of $\mathrm{CR}(\delta)$, defined by (1), from which the Pearson's chi-squared statistic and its companion measures of association can be derived in a straightforward manner. As an ANOVA-like partition, we consider the classical definition of inner products and orthogonality conditions for partitioning a measure of association belonging to the space $\Re^{I\times J\times K}$; for more details see [5, 11] and [2, Chapter 11]. Therefore, the general partition of (1) can be written as

$$\begin{aligned}\mathrm{CR}(\delta)=&\frac{2n}{\delta(\delta+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij\bullet}\left\{\left(\frac{p_{ij\bullet}}{p_{i\bullet\bullet}p_{\bullet j\bullet}}\right)^{\delta}-1\right\}\\&+\frac{2n}{\delta(\delta+1)}\sum_{i=1}^{I}\sum_{k=1}^{K}p_{i\bullet k}\left\{\left(\frac{p_{i\bullet k}}{p_{i\bullet\bullet}p_{\bullet\bullet k}}\right)^{\delta}-1\right\}\\&+\frac{2n}{\delta(\delta+1)}\sum_{j=1}^{J}\sum_{k=1}^{K}p_{\bullet jk}\left\{\left(\frac{p_{\bullet jk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{\delta}-1\right\}\end{aligned}$$

$$+ \frac{2n}{\delta(\delta+1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} p_{ijk} \left\{ \left(\frac{p_{ijk}}{\alpha p_{\bullet j \bullet} p_{\bullet\bullet k}} \right)^{\delta} - 1 \right\}$$
$$= \mathrm{CR}_{IJ}(\delta) + \mathrm{CR}_{IK}(\delta) + \mathrm{CR}_{JK}(\delta) + \mathrm{CR}_{IJK}(\delta). \quad (5)$$

Here we can see that there are four terms in the partition. The first three terms, $\mathrm{CR}_{IJ}(\delta)$, $\mathrm{CR}_{IK}(\delta)$ and $\mathrm{CR}_{JK}(\delta)$, are the bivariate Cressie-Read divergence statistics that assess the row-column, row-tube and column-tube association, respectively. Therefore, these measures are asymptotically chi-squared random variables with $(I-1)(J-1)$, $(I-1)(K-1)$ and $(J-1)(K-1)$ degrees of freedom, respectively. The last term, CR_{IJK}, is the measure of three-way, or *trivariate*, association between all three variables and is asymptotically a chi-squared random variable with $(I-1)(J-1)(K-1)$ degrees of freedom.

2 Application

In recent years, sustainable development goals (SDG) have become increasingly important. Decision makers everywhere need data and statistics that are accurate, timely, sufficiently disaggregated, relevant, accessible and easy to use. Here among the plethora of authoritative SDG data sources (available at https://unstats-undesa.opendata.arcgis.com/), we study the association between the *Renewable energy share* in the *total final energy consumption indicator* (RES), the *indicator of adjusted emission growth rate for black carbon indicator* (BCA) and the *Geographical area* (Africa, America, Asia, Australia, Caribbean islands, Europe) of 186 countries in 2018 (GEO). These three variables (RES, BCA and GEO) can be cross-classified to produce the $4 \times 4 \times 6$ contingency of Table 1. We now describe the nature of the three variables and examine the partition of $\mathrm{CR}(\delta)$ – see (5) – for the table where $\delta = 1, 1/2$ and $2/3$.

The four categories of the row variable RES are (0, 34.5], (34.5, 53.7], (53.7, 80.7] and (80.7, 100]. The column variable *BCA* also has four ordered categories; (0, 8.13], (8.13, 23.1], (23.1, 49.7] and (49.7, 96.4]. These variables are formed by dividing each continuous variable into quantiles. While both are ordinal variables we shall be treating them as nominal. The categories of the tube variable GEO are the six areas mentioned above.

We study the association of these three variables through the partition of (1); see (5). Since Table 1 has many zero cell frequencies, we compare the results of the partition of Pearson's chi-squared statistic with the results obtained from the partition of the three-way versions of the Freeman-Tukey statistic (T^2) and the Cressie-Read statistic (CR). We consider these last two statistics since there is a strong presence of overdispersion in data; see [4].

Pearson's chi-squared of Table 1 is $\mathrm{CR}(1) = 290.035$. With 84 degrees of freedom, this statistic has a p-value that is less than 0.001 and so a statistically significant association exists between the three variables. Table 2 summarises the partition

of this statistic which can be obtained by considering the three-way Cressie-Read divergence statistic with $\delta = 1$. All four terms of the partition are statistically significant with a p-value less than 0.001 except for the trivariate association term whose p-value is 0.002.

We take into account the presence of sparse data by investigating the partition of the Freeman-Tukey statistic ($\delta = 1/2$) and of the Cressie-Read statistic ($\delta = 2/3$). These two statistics are $\mathrm{CR}(1/2) = 250.491$ and $\mathrm{CR}(2/3) = 259.305$, respectively, and like $\mathrm{CR}(1) = X^2$ each has a p-value that is less than 0.001 thereby confirming that there exists a statistically significant association between the variables. Table 2 provides a summary of the partition of T^2 and CR and shows, like X^2, that each of the bivariate terms have a p-value that is less than 0.001. However, the trivariate association term is no longer statistically significant; see Table 2. This is likely due to the sparsity of many of the cell frequencies. Thus we can make the following

Table 1 Cross-classification of RES, BCA and GEO

	BCA			
RES	$(0, 8.13]$	$(8.13, 23.1]$	$(23.1, 49.7]$	$(49.7, 96.4]$
		Africa		
(0, 34.5]	1	0	4	8
(34.5, 53.7]	0	2	4	19
(53.7, 80.7]	3	2	1	4
(80.7, 100]	1	1	0	1
		America		
(0, 34.5]	1	0	6	2
(34.5, 53.7]	0	2	0	1
(53.7, 80.7]	0	3	1	0
(80.7, 100]	1	3	2	1
		Asia		
(0, 34.5]	5	1	2	0
(34.5, 53.7]	6	2	3	1
(53.7, 80.7]	5	3	3	3
(80.7, 100]	4	2	0	0
		Australia		
(0, 34.5]	1	0	0	0
(34.5, 53.7]	0	1	2	0
(53.7, 80.7]	1	1	5	0
(80.7, 100]	1	0	0	0
		Carribean		
(0, 34.5]	3	2	0	0
(34.5, 53.7]	1	1	0	1
(53.7, 80.7]	0	0	0	0
(80.7, 100]	2	3	0	0
		Europe		
(0, 34.5]	1	0	0	0
(34.5, 53.7]	0	0	1	0
(53.7, 80.7]	0	7	5	0
(80.7, 100]	3	11	7	3

Table 2 Partition of $\mathrm{CR}(\delta)$

Association	Term	%	*df*	*p*-value
	$\mathrm{CR}(1) = X^2$			
$I \times J$	34.970	12%	9	<0.001
$I \times K$	82.816	29%	15	<0.001
$J \times K$	95.677	33%	15	<0.001
$I \times J \times K$	76.573	26%	45	0.002
X^2	290.035	100%	84	<0.001
	$\mathrm{CR}(1/2) = T^2$			
$I \times J$	35.112	14%	9	<0.001
$I \times K$	83.340	33%	15	<0.001
$J \times K$	92.4207	37%	15	<0.001
$I \times J \times K$	39.619	16%	45	0.699
T^2	250.491	100%	84	<0.001
	$\mathrm{CR}(2/3) = CR$			
$I \times J$	34.988	13%	9	<0.001
$I \times K$	82.653	32%	15	<0.001
$J \times K$	93.105	36%	15	<0.001
$I \times J \times K$	48.559	19%	45	0.332
CR	259.305	100%	84	<0.001

conclusions about the nature of the association between the variables of Table 1. There is a statistically significant association between

- the total final energy consumption and the indicator of adjusted emission growth rate for black carbon,
- total final energy consumption and the geographical area,
- the indicator of adjusted emission growth rate for black carbon and the geographical area,

while no such association exists between all three variables.

References

1. Agresti, A.: Categorical Data Analysis (2nd ed). Wiley, New York (2002)
2. Beh, E. J., Lombardo, R.: Correspondence Analysis: Theory, Practice and New Strategies. Wiley, Chichester (2014)
3. Beh, E. J.: Simple correspondence analysis: A bibliographic review. International Statistical Review **72**, 257–284 (2004)
4. Beh, E. J., Lombardo, R., Alberti, G.: Correspondence analysis and the Freeman-Tukey statistic: A study of archaeological data. Computational Statistics and Data Analysis **128**, 73–86 (2018)
5. Carlier, A., Kroonenberg, P. M.: Biplots and decompositions in two-way and three-way correspondence analysis. Psychometrika **61**, 355–373 (1996)

6. Cressie, N. A. C., Read, T. R. C.: Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B **46**, 440–464 (1984)
7. Freeman, M. F., Tukey, J. W.: Multinomial goodness-of-fit tests. The Annals of Mathematical Statistics **21**, 607–611 (1950)
8. Greenacre, M.: Power transformations in correspondence analysis. Computational Statistics and Data Analysis **53**, 3107–3116 (2009)
9. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
10. Lancaster, H. O.: Complex contingency tables treated by the partition of chi-square. Journal of of the Royal Statistical Society B **13**, 242–249 (1951)
11. Lombardo, R., Takane, Y., Beh, E. J.: Familywise decompositions of Pearson's chi-square statistic in the analysis of contingency tables. Advances in Data Analysis and Classification **14(3)**, 629–649 (2019)
12. Neyman, J.: Contribution to the theory of certain test criteria. Bulletin de L'Institut International de Statistique **24**, 44–86 (1940)
13. Neyman, J.: Contributions to the theory of the χ^2 test. In: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability , 239–273 (1949)
14. Pardo, M. C.: An empirical investigation of Cressie and Read tests for the hypothesis of independence in three-way contingency tables. Kybernetika **32**, 175–183 (1996)
15. Pardo, L., Pardo, M. C.: Minimum power-divergence estimator in three-way contingency tables. Journal of Statistical Computation and Simulation **73**, 819–831 (2003)
16. Pardo, J. A.: An approach to multiway contingency tables based on ϕ-divergence test statistics. Journal of Multivariate Analysis **101**, 2305–2319 (2010)

Analysis of maximum precipitation in Thailand using non-stationary extreme value models

Thanawan Prahadchai, Juyoung Hong, Piyapatr Busababodhin, and Jeong-Soo Park

Abstract To investigate any changes in the magnitude or scale of maximum precipitation, in this study, we have built non-stationary models for annual maximum daily (AMP1) and 2-days rainfall (AMP2) data observed between 1984-2020 years by 71 stations and between 1960-2020 by 8 stations over Thailand. The generalized extreme value (GEV) distributions are used to model these data. Various time dependent of GEV model are considered, which consisted of totally 16 candidates. On each station, best model is selected by using two information criteria (Bayesian and Akaike information criteria; BIC, AIC) among these candidates. After a model is built, the return levels corresponding some years are calculated and predicted to the future. We found some evidence of increasing (decreasing) trends in AMP1 for 10 (4) stations in Thailand using BIC.

Key words: Bootstrap, Gumbel distribution, Heavy rainfall, Maximum likelihood estimation

1 Introduction

Extreme precipitation can result in floods and landslides, accompanied with a loss of life and costly damage of infrastructure. Thus, understanding and projecting heavy rainfall is of significant importance to climate change impact, adaptation, and vulnerability assessments.

[1] Thanawan Prahadchai, Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea; email: tanawanp.st@gmail.com

Juyoung Hong, Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea; email: hjy_stat@naver.com

Piyapatr Busababodhin, Department of Mathematics, Mahasarakham University, Mahasarakham 44150, Thailand; email: piyapatr.b@msu.ac

Jeong-Soo Park, Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea; email: jspark@jnu.ac.kr

Non-stationarity in heavy rainfall time series is often apparent in the form of trends because of long-term climate changes. Park et al.[4] studied the NS GEV distribution in modelling extreme precipitation over South Korea. Some researchers have modeled heavy rainfalls in Thailand. Busababodhin et al. [1] and Khongthip et al. [3] studied on extreme rainfall in Northeast and Upper North Thailand, respectively, using 3 NS GEV models. In this study, we present an application of NS GEV distributions to model extreme rainfall data throughout Thailand.

2 Data and Climatology

Figure 1 depicts a map of Indochinese peninsula including the Thailand showing observation stations and climatic regions. Thailand is located in the tropical area between latitudes 5°37′ N to 20°37′ N and longitudes 97°22′ E to 105°37′ E.

The original data of daily precipitation consist of measurements from 00h to 24h throughout the day. The data records from 1984 to 2020 by 71 stations and from 1960 to 2020 by 8 stations in Thailand. The data is available from the Thai Meteorological Department (TMD) [6]. The TMD classified 6 climatic regions as follows with number of stations in parenthesis: North (20), Northeast (17), Central (13), Eastern (10), Southeast (13), and Southwest (6). These regions are depicted in Figure 1. In the Southwest region receives a lot of rain and reaches a peak in September. On the contrary, there is a lot of rain in the Southeast region, whose peak in November remained until January of the following year.

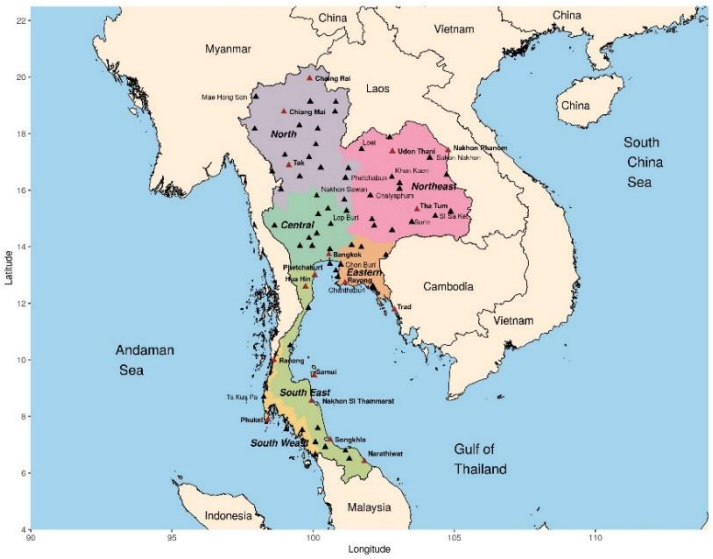


Figure 1: Map of Indochinese peninsula including the Thailand showing 79 observation stations and six climatic regions with different colors. The Thailand is located between latitudes 5°37′ N to 20°37′ N and longitudes 97°22′ E to 105°37′ E.

3 Methodology

3.1 Time Dependent Models for Extreme Value

The GEVD, which is asymptotically supportive and as flexible as the three well-known extreme value distributions, is widely used to analyse univariate extreme values [2]. The case for $\xi = 0$, is a Gumbel distribution (GD), whereas the cases $\xi > 0$ and $\xi < 0$ of GEVD are known as Fréchet and the negative Weibull distributions, respectively. The non-stationary models considered in this study are presented in Table 1. It consists of 16 models totally: 8 of GEVD and 8 of GD.

Table 1: Functional forms of parameters for time dependent non-stationary extreme value models.

Models	μ**(t),** σ**(t)**	***Parameters***	
		σ	ξ
M00	$\mu = \mu_0$	Constant	{Constant or 0}
M10	$\mu = \mu_0 + \mu_1 \times t$	Constant	{Constant or 0}
M20	$\mu = \mu_0 + \mu_1 \times t + \mu_2 \times t^2$	Constant	{Constant or 0}
M30	$\mu = \mu_0 + \mu_1 \times \exp(-\mu_2 \times t)$	Constant	{Constant or 0}
M01	$\mu = \mu_0$ $\sigma = \exp(\sigma_0 + \sigma_1 \times t)$		{Constant or 0}
M11	$\mu = \mu_0 + \mu_1 \times t$ $\sigma = \exp(\sigma_0 + \sigma_1 \times t)$		{Constant or 0}
M21	$\mu = \mu_0 + \mu_1 \times t + \mu_2 \times t^2$ $\sigma = \exp(\sigma_0 + \sigma_1 \times t)$		{Constant or 0}
M31	$\mu = \mu_0 + \mu_1 \times \exp(-\mu_2 \times t)$ $\sigma = \exp(\sigma_0 + \sigma_1 \times t)$		{Constant or 0}

3.2 Model Choice

We used maximum likelihood (ML) estimation method to estimate the parameters of Non-stationary (NS) GEV Models. This study selects the suitable model with the Akaike information criterion (AIC) and Bayesian information criterion (BIC) [5]. The model having smallest AIC or BIC is selected. After the best models are determined for each station, the next step is to derive the return levels which is the level exceeded on average only once in every T years [2].

4 Results

4.1 Selected Models

Table 2 presents parameter estimates and standard errors with in parenthesis computed from the non-stationary GEV models which were selected by the BIC for AMP1 data. The standard errors are obtained by parametric bootstrap technique.

Table 2: Non-stationary extreme value models selected based on the smallest Bayesian information criterion (BIC), parameter estimates, and standard errors in the parenthesis from the annual maximum daily rainfall (AMP1) data.

Name	***Model***	$\widehat{\mu}_0$	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_0$	$\widehat{\sigma}_1$	$\widehat{\xi}$
Chaiyaphum	GEV30	75.25 (2.70)	-50.76 (23.38)	0.44 (0.20)	12.76 (2.21)		0.33 (0.19)
Chiang Mai	GUM10	58.41 (5.07)	0.53 (0.14)		18.82 (1.97)		0
Phitsanulok	GUM01	82.78 (3.24)			3.49 (0.24)	-0.03 (0.01)	0
Ubon Ratchathani	GUM20	115.4 (10.3)	-1.45 (0.8)	0.03 (0.01)	26.27 (2.67)		0
Yala Agromet	GEV10	98.16 (9.98)	1.1 (0.39)		32.14 (5.48)		0.38 (0.17)
Nakhon Si Thammarat	GUM10	123.64 (19.73)	1.44 (0.56)		73.27 (7.70)		0

4.2 Return Levels

Figure 2 shows scatter plot of the AMP1 as time changes from the past to the future up to year 2040 for 6 locations. Lines over the plot represent the return levels corresponding to 2, 20, 50, 100, and 200 return periods, obtained from the selected NS GEV models. We can see some changes in magnitude as well as in scale of return levels.

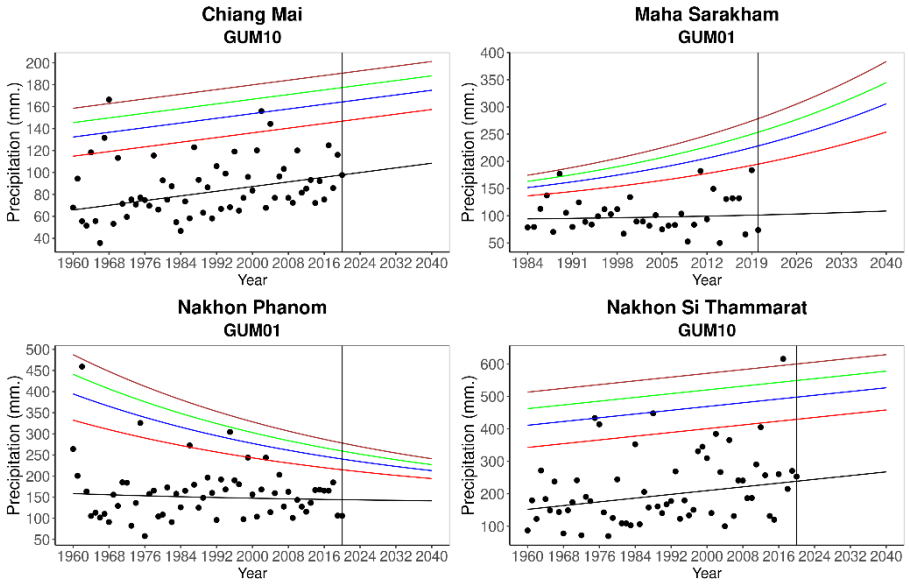


Figure 2 : Scatter plot of the annual maximum daily rainfall as time changes from the past to the future up to year 2040 for 6 locations with lines of return levels corresponding to 2, 20, 50, 100, and 200 return periods.

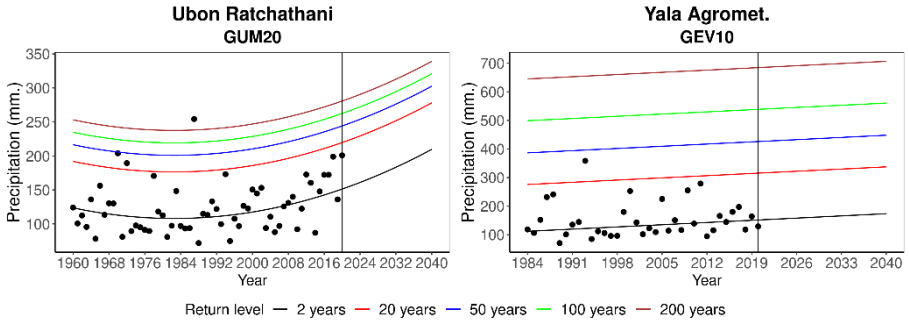


Figure 3 (*Continued*): Scatter plot of the annual maximum daily rainfall as time changes from the past to the future up to year 2040 for 6 locations with lines of return levels corresponding to 2, 20, 50, 100, and 200 return periods.

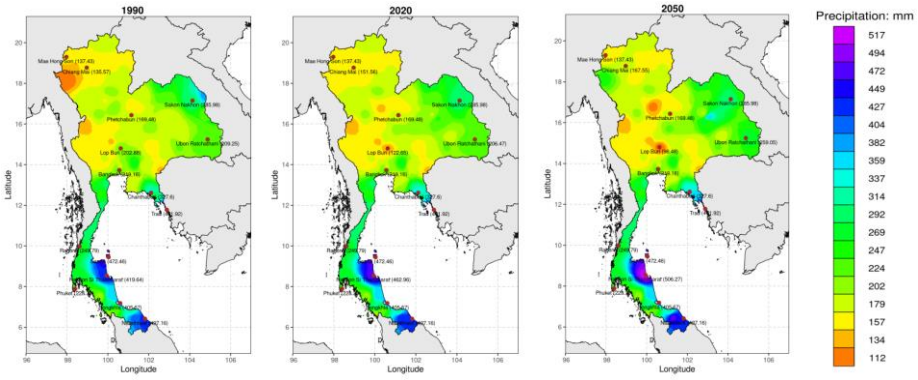


Figure 3: Maps of 50-year return levels of 1990, 2020, and 2050 estimated for the annual maximum daily precipitation data in Thailand.

5 Conclusions

In 79 stations, 32 (14) non-stationary GEV models for AMP1 data and non-stationary GEV 28 (12) models for AMP2 were selected based on AIC (BIC). The extreme rainfall in the northwest region including Mae Sa Riang and in the northeast region including Maha Sarakham are increasing. Whereas maximum precipitation in the central region including Lop Buri and Phitsanulok and in Nakon Phanom in the northeast region are decreasing. Downpour in the southeast region including Nakon Si Tammarat and Narathiwat are increasing with very heavy precipitation as Figure 3.

Funding

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No.2020R1I1A3069260) and BK21 FOUR

(Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). Piyapatr's work was supported and funded by Mahasarakham Univeristy, Thailand.

Acknowledgments

Observational data in the Thailand are provided by TMD at http://www.tmd.go.th/. We thank all contributors to the numerical R packages which were crucial for this work. The authors are grateful to Yire Shin who provided valuable comments to improve this paper.

References

1. Busababodhin, P., Siriboonand, M., Kaewmun, A.: Modeling of Extreme Precipitation in Upper Northeast of Thailand. Burapha Science Journal (in Thai) **20**, 106–117 (2015)
2. Coles, S.: An introduction to statistical modeling of extreme values, London: Springer (2001) doi: https://doi.org/10.1007/978-1-4471-3675-0
3. Khongthip, P., Khamkong, M., Bookamana, P.: Modeling Annual Extreme Precipitation in upper Northern Region of Thailand. Burapha Science Journal (in Thai) **18**, 95–104 (2013)
4. Park, J.S., Kang, H.S., Lee, Y.S., Kim, M.K.: Changes in the extreme daily rainfall in South Korea. International Journal of Climatology **31**, 2290–2299 (2011) doi: https://doi.org/10.1002/joc.2236
5. Schwarz, G.: Estimating the dimension of a model. The annals of statistics, pp. 461–464 (1978)
6. Thai Meteorological Department.: General Climatic Conditions (2021) Available online: https://www.tmd.go.th. Cited 17 May 2021

Assessing citizens' participation to urban transformation: a review of quantitative methods

Valutare la partecipazione dei cittadini alla trasformazione urbana: una review dei metodi quantitativi

Tregua, M. and Scaglione, M.

Abstract This research deals with citizens participation to the transformation of local contexts to identify the ways scholars investigated this phenomenon through quantitative methods. A systematic review was run on Web of Science, leading to select 948 contributions assessing citizens participation through quantitative methods. The analysis of these contributions led to highlight a growing trend in the use of quantitative methods, the opportunities brought by multiple data sources and longitudinal datasets, as well as a struggle between general analyses and observations limited to one specific city issue affected by citizen participation.

Abstract *Questa ricerca analizza gli approcci con cui gli studiosi hanno osservato tramite metodi quantitativi la partecipazione dei cittadini alla trasformazione urbana. Una review sistematica su 948 lavori ha mostrato un trend in netta crescita per analisi quantitative, l'esistenza di opportunità dovute alla presenza di più fonti di dati e di dataset longitudinali, così come una tensione tra analisi di carattere generale e analisi specifiche su singoli temi con riguardo alla partecipazione dei cittadini al processo di trasformazione.*

Key words: citizen participation, smart city, citizen engagement, quantitative methods.

1 Introduction

[1] Marco Tregua, University of Naples Federico II; email: marco.tregua@unina.it
Marco Scaglione, SAVA srl; email: scaglione@studio-visintin.it

Citizens participation is a key topic in cities' transformation as stated in the Leipzig Charter in 2007[1], when defining the support the City Labs could have offered to cities undergoing a process of change. Among the different prescriptions of the Charter, one is directly addressed to citizens, as Article 10 states that: "implementation-oriented planning tools should: be coordinated at local and city-regional level and involve the citizens and the agents that can contribute to shape the future economic, social and environmental quality of territories".

Scholars, too, focused on the contribution citizens may offer to cities, mainly when transformation processes are in progress; indeed, participation to democracy started being theorized in last century (e.g., [3]), also to understand how the intervention of citizens may divest politicians as well as urban specialists – city managers, city innovation managers, and so on. The attention paid by scholars to citizens participation increased along time, especially when the notion of smart city started inspiring new studies; indeed, since last decade the conceptualizations and projects trying to make cities smart – or even smarter – developed in several fields of science and in most of the areas of the world. Indeed, scholars identified citizens participation as a key element to pursue social and sustainable goals ([18]), as a relevant support in the policy-making process ([11]) due to the direct knowledge and the full awareness of local needs and interventions to be done, and as a factor leading to a clear theoretical and empirical distinction [7] between technology-centred cities transformation (e.g., digital city or cyber city) and human-centred cities transformation (i.e., smart city). Additionally, citizens participation is basically investigated in two ways, namely according to the physical or digital participation ([13]), and the individual or collective participation to decision-making process in cities ([9]). Due to such an attention, the topic has been framed and investigated in different ways and very often a local focus has been adopted to grasp meaning from empirical evidence of contexts in which people played a key role in supporting a city transformation (e.g., [2, 27]). Anyhow, the measurement of citizens participation still shapes an open question, since there are multiple methods as well as several factors to be considered in setting a proper tool for measuring the level of participation as well as its effects. Therefore, this research aims at reviewing the quantitative methods used by scholars in assessing the participations of citizens to the transformation processes of their cities. Consequently, the remainder of this paper introduces the criteria used in setting this review of methods, then a critical analysis of these methods is presented, leading to consider opportunities emerging from the methods available as well as future research avenues.

2 Criteria for literature review

[1] https://urbact.eu/leipzig-charter

To collect the extant contributions on citizens participation, first of all we structured queries in scholarly collections; therefore, we identified citizens participation, citizens engagement, and citizens involvement as the three topics shaping the debate about the contribution citizens may offer to the local development of urban policies. Besides being aware of the differences among the three definitions, we preferred to consider all of the contributions centred on at least one of these definitions to have a wider view. After selecting the topics of the queries, we decided to launch them – in both their singular and plural form – in Web of Science – Web of Knowledge, Core Collection, as it is the most common way to run such an analysis [26]. Through these queries we achieved a list of 8,220 contributions, then we filtered them according to the further following criteria:

- Document types: articles and book chapters, while other types of contributions were discarded due to their being most likely available in other forms
- Languages of publication: English

After applying the two above criteria we landed on a list of 5,093 contributions; then, we started analysing the abstract of these contributions – thus, we discarded 471 contributions without an abstract –to select the ones that:

- consider cities transformation (e.g., smart city) and discarded the ones dealing with other topics as disaster recovery, taxation, and issues not representing a transformation itself;
- use quantitative methods, thus we discarded conceptual/theoretical studies or contributions adopting qualitative methods or con
- adopt the perspective of evaluating citizens participation, thus we put aside the contributions focused on government strategies to involve citizens as well as the ones just addressing tools to have citizens actively participating to the decision-making process.

Finally, our dataset consists of 948 entries out of 4,622.

3 Review and classification of methods

The analysis of citizenship participation has been traditionally constrained by the level of shared decision allowed and by the purposes of the engagement policy pursued by the authority. This endowed the analyst only with a little amount of data for administrative units, mostly in the form of Citizen Surveys or interviews performed by the researchers themselves. This is the reason behind the great success that path analysis methods, such as SEM, MANOVA and LDA, have achieved in our dataset (approx. 61%). A first innovation in this branch of literature has been provided by Matsui et al. [21], which is the first analysis in our dataset to combine households surveys and administrative data and builds a predictive model of recycling behaviour. The paper by Chen et al. [5] is the first entry in our collection to rely on *new data* analysis to study the Taipei Citizen Complaints System, which

involved letters and faxes, monitoring of the number phone calls, visit in person to the city bureaus, letter to newspapers and a classification of the emails received by the mayor at his institutional inbox. Another strategical innovation is the one introduced by Ripat et al. [23], that helped the Canadian Public Works Department to gather information on community participation and winter mobility, siding the usual interviews methodologies with the development of a proper walking log to gather further information on the walking patterns, sidewalk walkability and other issues related to walking in winter.

Another branch of literature [15, 24, 28] focus on secondary survey analysis for comparing the outcome of different participation experience in different countries and the driving factors of this outcomes. In this studies SEM is used to build indexes representing different aspects of the experience, indexes representing different exogenous variables which are supposed to affect the outcomes and standard regression techniques are then used to provide the sensibility of the different outcomes to the different exogenous variables.

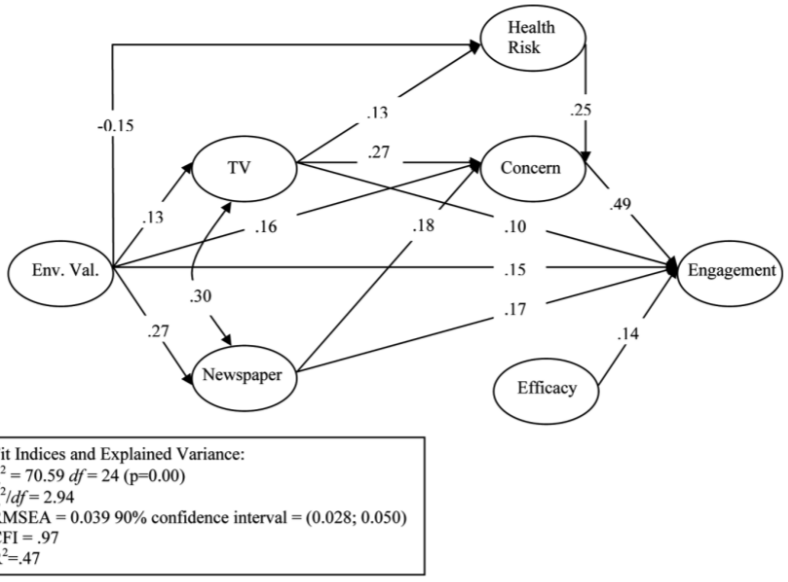


Figure 1: Representation of SEM output in Hart et al. (2011).

Damurski [8] is an example of the novelty introduced on the subject by the digitalization of the citizen-institution relationship, providing a classification of e-participation tools (mostly website) with different criteria, answering several questions related to their design and calculating scores on specific features, as spatiality, interactivity and transparency.

Methodologically Foster-Fishman [12] introduces a novelty, making a dynamic comparison that allowed them to investigate the processes promoting citizen participation, analysing the collection of 542 longitudinal surveys from the resident

of a small American Midwestern city implementing community change initiative. The presence of repeated measures allowed the researcher to test hypothesis on the changes before and after the program, through tests on SEM coefficients as showed in the Figure 2.

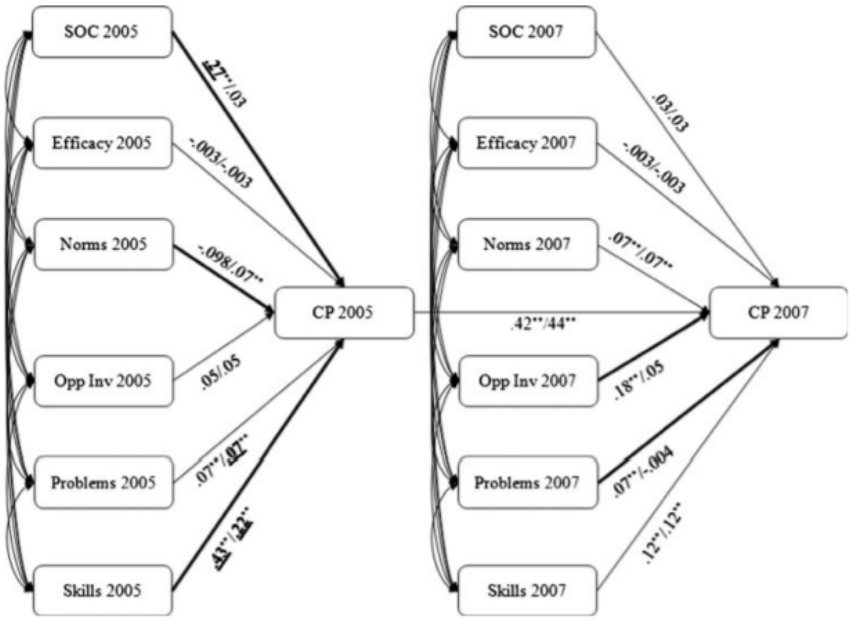


Figure 2: Final model in Foster-Fishman [12], bold and underlined estimates indicate a significant path difference between the two time points, for example the path between sense of community (SOC) and citizen participation (CP) is a significantly greater predictor in 2005 compared to 2007.

While the articles cited so far present analysis of all the data generated by a specific enquiry, Akinboade et al. [1] is the first standing example of a reasoned sampling procedure that allowed a meaningful analysis of variance between different group providing a clear statement on the satisfaction on service delivery by age, gender, and district of residence in South Africa. Cornwall and Shankland [6] and Guo and Neshkova [14] are further examples of solid stratified sampling strategies in this context.

The spreading of web technology and big data transformed the possibility to analyse the degree of involvement in the urban bottom-up movement, Niederer and Priester [22] combined content analysis and network analysis to piece together a longitudinal analysis of 14 years of bottom-up initiatives and analysing forty websites related to Amsterdam districts.

Zheng and Pan [29] applied LCA and multinomial regression analysis to determine the types and the socioeconomic drivers of online citizen participation in China. Maboudi and Nadi [20] analysed the 650,000 contributions emerged from the crowdsourcing initiative for Egyptian Constitution in 2012, an unprecedented

feedback loop of online public input and constitutional changes. They tested two random coefficient multilevel models to explain the changes in constitutional draft and the different kind of inputs and supports received in the crowdsourcing initiative, using text analysis tools to code the comments received for each draft. He et al. [17] examines the role of e-participation in urban China triangulating online questionnaires conducted in Bejing and media data analysis of 12 public protests that took place in several Chinese cities in 2011-2014. Fenoll Tomé and Cano-Orón [10] analysed citizen engagement in the comments written on the Facebook pages of the major Spanish political parties during the 2015 general election campaign, performing computerized content analysis of the comments to identifies words, emotional valence, and the type of language used in the interaction. Hayes and Lawless [16] also perform content analysis of more than 10,000 stories about US House campaigns in 2010 and 2014 to investigate how news environment influences citizen engagement. A similar purpose has been pursued by Siyam et al. [25] which analysed e-participation rate through a collection of 55,809 tweets over a period of one year from the Twitter account of a progressive government in the Arab world (Dubai). Cantador et al. [4] also provide content analysis but they retrieve information from Open Government Data and analyse 24,800 proposal and 86,100 comments on Decide Madrid an electronic participatory budgeting platform. Taipei participatory budget platform has been similarly analysed in Kuo et al. [19].

4 Implications and Conclusions

The analysis mirrored quite a mature stream of research, as multiple methods have been used from 2007 on.

First and foremost, the analysis performed led to consider that a ‘one size fits all’ solution can’t be foreseen, due to the variety of issues affecting cities’ transformations, as well as due to the features of each context. Secondly, there has been a significant evolution, mainly due to two aspects, viz. an increase in transformative interventions in cities and the availability of more data; indeed, this latter favoured the use of quantitative methods as SEM (e.g., [28]). Thirdly, the sources used to assess participation expanded throughout time, since municipalities – and other local agencies – got awareness on the relevance of these measures, and scholars benefited from that (e.g., [5]). Fourthly, longitudinal analyses started emerging, mirroring the reason of cities transformation processes and offering a significant support to the assessment of both changes and participation. Finally, some analyses were focused on a specific topic (e.g, [16]), while others adopted a more general posture ([17]).

To sum up, the analysis of quantitative methods led to foreseen opportunities to combine methods, sources, as well as to focus on both general and specific topics; according to most of scholars investigating how to assess citizens participation the

trade-off for future research is between new analyses combining methods and tools and multiple analysis run at specific levels. This is also a call for future research on this topic, as well to address the limitations of this study, namely the chance to spot a tendency to combine city issues with research methods and contexts with research methods.

References

1. Akinboade, O. A., Mokwena, M. P., & Kinfack, E. C., Understanding citizens' participation in service delivery protests in South Africa's Sedibeng district municipality. *International Journal of Social Economics* **40**(5), 458-478, (2013).
2. Berntzen, L., & Johannessen, M. R., The role of citizen participation in municipal smart city projects: Lessons learned from Norway. In Smarter as the new urban agenda (pp. 299-314). Springer, Cham., (2016).
3. Burke, E. M., Citizen participation strategies. Journal of the American Institute of Planners, **34**(5), 287-294, (1968).
4. Cantador, I., Cortés-Cediel, M. E., & Fernández, M., Exploiting Open Data to analyze discussion and controversy in online citizen participation. *Information Processing & Management*, ***57***(5), 102301, (2020).
5. Chen, D. Y., Huang, T. Y., & Hsaio, N., The management of citizen participation in taiwan: A case study of Taipei city government's citizen complaints system. *International Journal of Public Administration*, ***26***(5), 525-547, (2003).
6. Cornwall, A., & Shankland, A., Cultures of politics, spaces of power: contextualizing Brazilian experiences of participation. *Journal of Political Power*, ***6***(2), 309-333, (2013).
7. D'Auria, A., Tregua, M., & Vallejo-Martos, M. C., Modern conceptions of cities as smart and sustainable and their commonalities. Sustainability, **10**(8), 2642, (2018).
8. Damurski, L., E-participation in urban planning: Online tools for citizen engagement in Poland and in Germany. *International Journal of E-Planning Research (IJEPR)*, **1**(3), 40-67, (2012).
9. Datta, A., Blockchain Enabled Digital Government and Public Sector Services: A Survey. In Blockchain and the Public Sector (pp. 175-195). Springer, Cham., (2021).
10. Fenoll Tomé, F. V., & Cano Orón, L., Citizen engagement on Spanish political parties' Facebook pages: Analysis of the 2015 electoral campaign comments. *Communication & society*, *2017*, Vol. **30**, Num. 4, p. 131-145, (2017).
11. Fernandez-Anez, V., Fernández-Güell, J. M., & Giffinger, R., Smart City implementation and discourses: An integrated conceptual model. The case of Vienna. Cities, **78**, 4-16, (2018).
12. Foster-Fishman, P. G., Collins, C., & Pierce, S. J., An investigation of the dynamic processes promoting citizen participation. *American journal of community psychology*, ***51***(3-4), 492-509, (2013).
13. Gaffney, C., & Robertson, C., Smarter than smart: Rio de Janeiro's flawed emergence as a smart city. *Journal of Urban Technology*, **25**(3), 47-64, (2018).
14. Guo, H., & Neshkova, M. I., Citizen input in the budget process: When does it matter most?. *The American Review of Public Administration*, ***43***(3), 331-346, (2013).
15. Hart, P. S., Nisbet, E. C., & Shanahan, J. E., Environmental values and the social amplification of risk: An examination of how environmental values and media use influence predispositions for public engagement in wildlife management decision making. *Society and Natural Resources*, **24**(3), 276-291, (2011).
16. Hayes, D., & Lawless, J. L., The decline of local news and its effects: New evidence from longitudinal data. *The Journal of Politics*, **80**(1), 332-336,q (2018).
17. He, G., Boas, I., Mol, A. P., & Lu, Y., E-participation for environmental sustainability in transitional urban China. *Sustainability Science*, **12**(2), 187-202, (2017).
18. Hollands, R. G., Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?. *City*, **12**(3), 303-320, (2008).

19. Kuo, N. L., Chen, T. Y., & Su, T. T., A new tool for urban governance or just rhetoric? The case of participatory budgeting in Taipei City. *Australian Journal of Social Issues*, ***55***(2), 125-140, (2020).
20. Maboudi, T., & Nadi, G. P., Crowdsourcing the Egyptian constitution: social media, elites, and the populace. *Political Research Quarterly*, ***69***(4), 716-731, (2016).
21. Matsui, Y., Tanaka, M., & Ohsako, M., Study of the effect of political measures on the citizen participation rate in recycling and on the environmental load reduction. *Waste Management*, ***27***(8), S9-S20, (2007).
22. Niederer, S., & Priester, R., Smart citizens: Exploring the tools of the urban bottom-up movement. *Computer Supported Cooperative Work (CSCW)*, ***25***(2-3), 137-152, (2016).
23. Ripat, J. D., Redmond, J. D., & Grabowecky, B. R., The winter walkability project: occupational therapists' role in promoting citizen engagement. *Canadian Journal of Occupational Therapy*, ***77***(1), 7-14, (2010).
24. Royo, S., Yetano, A., & Acerete, B., Citizen participation in German and Spanish local governments: A comparative study. *International Journal of Public Administration*, ***34***(3), 139-150, (2011).
25. Siyam, N., Alqaryouti, O., & Abdallah, S., Mining government tweets to identify and predict citizens engagement. *Technology in Society*, ***60***, 101211, (2020).
26. Wang, X., Fang, Z., & Sun, X., Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. *Scientometrics*, **109**(2), 917-926, (2016).
27. Willems, J., Van den Bergh, J., & Viaene, S., Smart city projects and citizen participation: The case of London. In Public sector management in a globalized world (pp. 249-266). Springer Gabler, Wiesbaden, (2017).
28. Yetano, A., Royo, S., & Acerete, B., What is driving the increasing presence of citizen participation initiatives?. *Environment and Planning C: Government and Policy*, ***28***(5), 783-802, (2010).
29. Zheng, J., & Pan, Z., Differential modes of engagement in the Internet era: a latent class analysis of citizen participation and its stratification in China. *Asian Journal of Communication*, ***26***(2), 95-113, (2016).

Session of solicited contributes SS20 – *Statistical methods for health and environmental impact assessment*
Organizer and Chair: Fabrizio Maturo

Density estimation via Functional Data Analysis

Stima delle densità attraverso l'analisi dei dati funzionali

Stefano Antonio Gattone and Tonio Di Battista

Abstract Recent technological advances have eased the collection of big amounts of data in many research field. In this scenario an useful statistical technique is density estimation which represents an important source of information. One dimensional density functions represent a special case of functional data subject to the constraints to be non-negative and with a constant integral equal to one. Because of these constraints, densities functions do not form a vector space and a naive application of functional data analysis (FDA) methods may lead to non valid estimates. To solve this problem, by means of a suitable transformation densities are embedded in the Hilbert space of square integrable functions where standard FDA methodologies can be applied.

Abstract *I recenti sviluppi tecnologici permettono di raccogliere con facilità grandi quantità di dati in molti ambiti di ricerca. In questo scenario, una tecnica statistica utile è la stima della funzione di densità che rappresenta un'importante risorsa di informazioni. Le funzioni di densità unidimensionali rappresentano un caso speciale di dato funzionale soggetto ai vincoli di non negatività e con integrale costante e pari a uno. A causa di questi vincoli, le funzioni di densità non formano uno spazio vettoriale e un'applicazione diretta dei metodi di analisi funzionale può condurre a risultati non validi. Per risolvere il problema, le funzioni di densità sono immerse nello spazio di Hilbert dove possono trovare applicazione le tecniche standard dell'analisi funzionale.*

Key words: Constrained estimator, Functional Data Analysis, Probability density functions.

Stefano Antonio Gattone
DISFIPEQ, G. d'Annunzio University, Pescara, Italy, e-mail: gattone@unich.it

Tonio Di Battista
DISFIPEQ, G. d'Annunzio University, Pescara, Italy, e-mail: dibattis@unich.it

1 Introduction

Probability density functions (pdfs) can provide useful information on large-scale database since they provide more information than single summaries statistics such mean, variance, skewness and so on. They may also be a starting point for further analysis in non-parametric prediction models.

Pdfs represent a special case of functional data [10] since they must satisfy the constraints of being non-negative everywhere and present a constant integral constraint equal to one. These characteristics pose the pdfs in convex but non linear space where common FDA methods cannot be naively applied. See, for example, for functional principal components analysis of density functions the works in [2, 7]. To address this issue two main strategies can be found in the literature. In the first, the pdfs are mapped into a linear functional space through a suitably chosen transformation. Established methods for Hilbert space valued data can be applied to the transformed functions and the results are moved back into the density space by means of the inverse transformation. Typical transformations are the log [2] and the log hazard transformations [7]. In the second strategy, pdfs are treated as an infinite dimensional compositional data since they are part of some whole which only carry relative information. An approach based on compositional data methods has been sketched in [2], applying theoretical results in [4], which define a Hilbert structure on the space of densities. Similarly, [6] accounted for the specific characteristics of density functions in Bayes linear spaces, which results from the generalization to the infinite dimensional setting of the Aitchison geometry for compositional data [1]. The authors proposed the use of the centred log-ratio transformation, which represents an isometric isomorphism between the Bayes space of pdfs and the space of square-integrable real measurable functions.

This work deals with the estimation of the probability density function. Within this framework, two problems must be faced. The first relates to the use of standard FDA techniques when dealing with density functions. With this respect, a transformation-based approach is followed where density functions are defined as solutions of differential equations. Details of the proposed transformation are provided in Section 2. The second problem relates to the estimation of the bias and the variance of the estimator. A procedure based on delta method is given in Section 3.

2 Density functions as constrained functional data

Let $f(x)$ be a random probability density function whose support is a finite interval $\mathscr{I} = [a,b]$, and $\mathscr{F}(\mathscr{I})$ be the function space of pdfs on $\mathscr{I}$ such that $f(x) \geq 0$ and $\int_{\mathscr{I}} f(x)dx = 1$. The constrained estimation of the underlying function is redefined into an unconstrained one by using a differential equation method [9, 5]:

$$Df(x) = w(x)f(x) \tag{1}$$

with solution equal to

$$f(x) = C \exp \int_a^x w(u)du \tag{2}$$

where $C = [\int_{\mathscr{I}} \exp \int_{\mathscr{I}} w(u)du]^{-1}$ is a normalizing constant so that $\int_{\mathscr{I}} f(x)dx = 1$. The crucial point is that $w(u)$, $u \in [a,b]$ is a square integrable function free of constraints. The advantage of this transformation is that a constrained problem is changed to an unconstrained one. Let $w(x) = c^T \Phi(x)$ where Φ is a set of basis functions and c a vector of coefficients defining the basis expansion of $w(x)$. The density estimator is obtained by minimizing a particular metric in the space of density functions. For example, by using the Hellinger distance, the following criterion is minimized:

$$L(c) = \frac{1}{2} \int_{\mathscr{I}} \left(\sqrt{g(x)} - \sqrt{f(x;c)} \right)^2 dx. \tag{3}$$

Other distances can also be used: for instance the L_1 distance or the symmetrized version of the Kullback-Leibler divergence [3].

3 Functional estimation of density functions

In real applications, one has at hand an *iid* sample of data generated by each random density $f_i(x)$, $i = 1,\ldots,n$. Thus, there are two sources of random variation, the first generating the sample of densities and the second generating the samples of data [7]. In this work we will assume that a sample of n *iid* density functions $f_1(x), f_2(x), \ldots, f_n(x)$ is available.

Since the space of densities is convex, the mean function f_μ of the density process can be estimated by $\bar{f} = \frac{1}{n}\sum_{i=1}^n f_i(x)$. For a different method based on quantile synchronization which takes into account the horizontal variation in the densities see [12]. However, if one is interested in exploring the variability of the densities, for example by computing simultaneous confidence bands, even if the results may provide a good approximations they may not satisfy the characteristics of a density. This is the drawback of applying methods suitable for functions in the Hilbert space to functions that are in a non linear space. A naive solution could be to take the positive part and re-normalize.

Using transformation in equation (2) the sample of density functions $\{f_i(x)\}_{i=1}^n$ is transformed to the unconstrained functions $\{w_i(x)\}_{i=1}^n$. The mean and the covariance function of the sample $\{w_i(x)\}_{i=1}^n$ can be easily estimated using standard FDA tools and denoted with $\bar{w}(x)$ and $\gamma(x,x')$ for $x,x' \in \mathscr{I}$, respectively. The proposal is to work with the functions $\{w_i(x)\}_{i=1}^n$ where all the FDA tools can be coherently applied and then mapping back to the density space by putting the results into equation (2). It becomes necessary, then, to check if the transformation proposed is useful in practice. This is the goal of the extend version of this work.

References

1. Aitchison, J.: The statsitical analysis of compositional data. Journal of the Royal Statistical Society, Series B **44**, 139–177 (1982)
2. Delicado, P.: Dimensionality reduction when data are density functions. Computational Statistics & Data Analysis **55**, 401–420 (2011)
3. Devroye, L. and Gyorfi, L.: Nonparametric Density Estimation: The L_1 View. Wiley, New York (1985)
4. Egozcue, J., Diaz-Barrero, J., and Pawlowsky-Glahn, V.: Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica, Englis Series **22**, 1175–1182 (2006)
5. Gattone, S.A., Di Battista, T.: A functional approach to diversity profiles. Journal of the Royal Statistical Society, Series C **58**, 267–284 (2009)
6. Hron, K., Menafoglio, M., Templ, M., Hruzova, K. and Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics & Data Analysis **94**, 330–350 (2016)
7. Petersen, A. and Muller, H.: Functional data analysis for density functions by transformation to a Hilbert space. The Annals of Statistics **32**(1), 183–218 (2016)
8. Ramsay, J.O.: Monotone regression splines in action. Statistical science **1**, 425–441 (1988)
9. Ramsay, J.O.: Estimating smooth monotone functions. Journal of the Royal Statistical Society **60**, 365–375 (1998)
10. Ramsay, J.O. and Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, New York (2005)
11. Sen, R. and Ma, C.: Forecasting density function: an application in finance. Journal of Mathematical Finance **5**, 433–447 (2015)
12. Zhang, Z. and Muller, H.G.: Functional density synchronization. Computational Statistics & Data Analysis **55**, 2234–2249 (2011)

A new multivariate functional ANOVA approach for assessing air quality data amid COVID-19 pandemic

Un nuovo approccio funzionale dell'ANOVA multivariata per la valutazione dei dati sulla qualità dell'aria durante la pandemia COVID-19

Adelia Evangelista, Christian Acal, Ana M. Aguilera, Annalina Sarra, Tonio Di Battista and Sergio Palermi

Abstract To reduce the SARS-CoV-2 virus spreading, worldwide governments implemented a series of restriction measures that led to a downturn in several economic sectors. Recent studies have instead documented a respite to the environment. In this work, we evaluate the impact of lockdown on air quality of the urban area of Chieti-Pescara (Abruzzo region, Italy). To this end, we adopt a functional data analysis approach. Specifically, to check the differences between the temporal evolution of different pollutants (PM_{10}, $PM_{2.5}$, NO_2 and benzene) in terms of the location of measuring stations, a novel approach for multivariate FANOVA for independent measures, is proposed. The results obtained reveal changes in pollutants behaviour during the lockdown period.

Abstract *Per ridurre la diffusione del virus SARS-CoV-2, i governi di tutto il mondo hanno implementato una serie di misure restrittive che hanno portato a una recessione in diversi settori. Recenti studi hanno, invece, documentato una tregua per l'ambiente. In questo lavoro, valutiamo l'impatto del lockdown sulla qualità dell'aria dell'area urbana di Chieti-Pescara (regione Abruzzo, Italia). A tal fine, adottiamo un'analisi funzionale dei dati. Nello specifico, per verificare le differenze tra l'evoluzione temporale di diversi inquinanti (PM_{10}, $PM_{2.5}$, NO_2 e benzene) a seconda della posizione delle stazioni di monitoraggio, si propone un nuovo approccio basato sulla FANOVA multivariata per misure indipendenti. I risultati ottenuti rivelano cambiamenti nel comportamento di tutti gli inquinanti durante il periodo di lockdown.*

A. Evangelista, A. Sarra, T. Di Battista
Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. d'Annunzio, V.le Pindaro, 42; 65127 Pescara (Italy), e-mail: asarra@unich.it, e-mail: adelia.evangelista@unich.it, e-mail: tonio.dibattista@unich.it
C. Acal, A.M. Aguilera
Department of Statistics and O.R. and IEMath-GR, University of Granada (Spain), e-mail: chacal@ugr.es, e-mail: aaguilera@ugr.es
S. Palermi
Agency of Environmental Protection of Abruzzo (ARTA), V.le G. Marconi, 51; 65127 Pescara (Italy), e-mail: s.palermi@artaabruzzo.it

Key words: Air pollution, COVID-19/Lockdown, FDA/Functional Data Anlysis, Public health, Multivariate ANOVA

1 Introduction

Starting from the first outbreak, first identified in Wuhan (China), in late December 2019, the Coronavirus disease affected the entire world, posing major public health and governance concerns. Due to the wide spread of this pandemic disaster, authorities enforced different measures, resulting in prohibitions of various aspects of human activities. Global and local economy had intense damaging, especially in sectors such as tourism, commodity markets and transportation. On the other hand, several works have reported that highly industrialized zones of the world observed a remarkable reduction in air pollution, essentially due to restrictions placed upon industrial activities and the dropping in road transport. To this regard, [1] detected a significant reduction of air pollution in various areas of China and India; [2] studied the behaviour of the levels of pollutants across USA while [3] reported a sharp reduction in air pollution in large parts of Europe. Following these lines of research, in this paper, we investigate the effects of the quarantine policies adopted by the Italian Government on air quality in the urban area of Chieti-Pescara (Abruzzo region, Italy). By carrying out a functional data analysis (FDA), we compare the behaviour of different air pollutants in two different periods of time: before lockdown and during lockdown days. Starting from the foundations given by [4, 5], the FDA has been extensively used in the last decades, also in environmental studies, because FDA paradigm makes it possible to work with the entire time spectrum of pollutants time series, bringing additional information to be recovered from the data than in the vectorial approach (see, among others, [6], [7]). Within the methodological FDA framework, we proposed here a novel approach based on the Functional Analysis of Variance (FANOVA) for independent measures. The rest of the paper is organised as follows: Section 2 gives the description of the area under study and the air pollutants. Section 3 briefly defines the new methodology proposed while in Section 4 the main results obtained are illustrated. Finally, in Section 6 there are some concluding remarks.

2 Area of study

In this work, we take into account the metropolitan area of Chieti-Pescara (along the Adriatic coast of central Italy), defined of critical importance in terms of environmental pollution. Chieti-Pescara conurbation represents one of the most important industrial pole of the Abruzzo. In the last years, this area has registered an increase of industrial activity, as well as of the urban development. All these aspects make the metropolitan area the locus of growing environmental concerns for the high level

of resource consumption, greenhouse gas emissions and air quality pollution.

2.1 Air pollution data

Air pollution data for this analysis consists in hourly measurements of PM_{10}, $PM_{2.5}$, NO_2 and benzene, detected by the Regional Agency for the Environmental Protection (ARTA) of Abruzzo by means of five monitoring stations of the regional air quality network. The air quality monitoring sites of Pescara (Teatro d'Annunzio), Chieti and Francavilla are defined as *Urban Background type* (UB) due to their spatial location not affected by important pollution sources. On the other hand, the monitoring stations of Pescara (Via Firenze) and Montesilvano are named as *Urban Traffic type* (UB) because are nearby roadside, so they are influenced by traffic emissions. The pollutants data collected have been divided in two time intervals: the first is defined *pre-lockdown period* starting from the 1st February 2020 to the 10th of March 2020; the second is named *during-lockdown period* from the 11st of March to the 18th of April 2020.

Table 1 Net and % variation of pollutants concentration levels in the urban area of Chieti-Pescara

	UT		**UB**		
Net variation	**fi**	**mo**	**th**	**ch**	**fr**
NO_2	-13.9	-14.7	-21.2	-10.3	-7.6
PM_{10}	5.1	3.7	5.7	4.3	7.3
$PM_{2.5}$	2.9	2.2	3.1	4.4	4.1
Benzene	-0.31	-0.15	0.22	0.18	0.04
% variation					
NO_2	-57.9	-58.7	-65.2	-54.8	-49.1
PM_{10}	20.5	16.8	22.0	19.4	40.8
$PM_{2.5}$	19.0	15.6	19.8	26.7	34.4
Benzene	-32.57	-27.56	40.06	19.63	4.27

Acronyms of monitoring stations:
fi=Via Firenze; mo=Montesilvano; th=Teatro d'Annunzio; ch=Chieti; fr=Francavilla al Mare

An initial investigation of the net and percentage variations of the four pollutants in the different monitoring sites, before and during lockdown, is given in Table 1. The analysis shows that NO_2 recorded a significant and marked reduction both in traffic and in background monitoring stations. This is in line with our expectations, due to the collapse of vehicular traffic after the measures imposed by the government. The levels of particulate matter (PM_{10} and $PM_{2.5}$) registered an increment in all measuring sites. This could be reasonable knowing the peculiarity of this pollu-

tant. Finally, for benzene, we observed an opposite behaviour: for the UT stations the pollutant decreases, while it increases in the BT stations.

3 Methodological framework

In FDA approaches, the data analyzed are curves, or more typically functions, that varying over time, space, or other continuous support. The first approach proposed in this work has the aim of testing the equality, across two different conditions or periods, of mean functions related to a unique functional variable. Let $X_{jr}(t)$ be the sample functions, where $t \in T = [a,b]$ is the temporal time interval, $j = 1,...,n$ is the sample unit and $r = 1,...,R$ represents the number of the different periods or time (or conditions) to compare. In the current work, we compare only two different periods ($R = 2$). We define with $\mu_r(t) = E[X_{jr}(t)]$ the mean function associated to each functional variable in each condition or time period. The aim is to test $H_0 : \mu_1(t) = \mu_2(t)\ \forall t \in [a,b]$, against the alternative that its negation holds. The two statistics proposed to carry out the hypothesis testing are those introduced by [8] which take into account simultaneously the between and within variabilities. Smaga's statistics are defined as:

$$\mathscr{D}_n = n \int_T \frac{\left(\overline{X}_1(t) - \overline{X}_2(t)\right)^2}{\hat{K}(t,t)}\, dt,$$

$$\mathscr{E}_n = sup_{t\in[a,b]} \left\{ \frac{n\left(\overline{X}_1(t) - \overline{X}_2(t)\right)^2}{\hat{K}(t,t)} \right\},$$

where $\hat{K}(t,t) = \frac{\sum_{j=1}^{n}\left[(X_{j1}(t)-\overline{X}_1(t))-(X_{j2}(t)-\overline{X}_2(t))\right]^2}{n-1}$.

$\mathscr{D}_n$ and $\mathscr{E}_n$ can be computed by considering the basis expansion. A further stage in our analysis was to test the equality of the multivariate dimensional mean functions for independent groups (g). To comply with this aim, we consider the multivariate FANOVA for independent measures, introduced and described in detail in [9]. An important result about FANOVA has been obtained by [10] which demonstrate that FPCA of $X(t)$ is the same as to apply a Multivariate PCA on the matrix $A\Psi^{1/2}$. The hypothesis to test is:

$$\left\{ H_0 : \mu_1(t) = ... = \mu_g(t)\ \forall t \in [a,b].\right.$$

against the alternative that its negation holds. In this setting, two problems are encountered. First, the multivariate homogeneity tests do not perform well with high dimensional vectors and the number of basis functions needed for an accurate approximation of sample is usually high. As a solution, we propose to test the multivariate homogeneity on the vectors of the most explicative principal components scores. This new methodology proposes an extension of the novel parametric and

nonparametric approaches introduced using Functional Principal Component Analysis for univariate functional data ([11]) to the multivariate case.

4 Results

The impact of lockdown measures on air quality has been investigated with the functional testing procedures described in Sect. 3. To carry out the analysis, we convert the discrete values into curves, by means of cubic B-spline smoothing with 20 basis functions. Firstly, we adopted a FANOVA for repeated measures to statistically prove the results obtained in Table 1. The statistics $\mathscr{D}_n$ and $\mathscr{E}_n$ are used to test the within and between group variability and the p-values obtained by means of permutation tests. The results of the tests are significant for NO_2, PM_{10} and $PM_{2.5}$, suggesting that there are differences for these pollutants in the means curves before and during the lockdown period. Not statistical significance was found, instead, for benzene. However, considering that the results of hypotheses testing for benzene are very close to the limit region and taking into account the small sample size, we can also conclude that there are also differences in the means curves of benzene for the two time periods. A second step of our study consists in the multivariate analysis of variance for independent measures. We tested if there are differences between the temporal evolution of all pollutants in terms of the location of measuring station. According to the results displayed in in Table 2, significant differences were found in terms of the location of the monitoring stations in relation to PM_{10} (before lockdown) and benzene (during lockdown).

Table 2 Multivariate FANOVA for independent measures

p-value	**BL**	**DL**
All pollutants	0.000	0.302
NO_2	0.562	0.272
PM_{10}	0.000	0.306
$PM_{2.5}$	0.889	0.685
Benzene	0.186	0.000

Acronyms:
BL=Before Lockdown; DL=During Lockdown

5 Conclusion

In this work, changes in air pollution during the COVID-19 pandemic have been evaluated by means of a functional analysis of variance with univariate repeated

measures and multivariate independent measures. The results has proven a possible misclassification of air monitoring stations in the urban area of Chieti-Pescara, probably due to the NO_2. since the proposed technique failed to discriminate between UB and UT measuring sites, despite the fact that NO2, in urban areas, is a pollutant mostly produced by traffic emissions. This result is of great importance for environmental protection agencies which should identify the presence of redundant or misclassified monitoring sites, in order to reduce the cost of pollution monitoring and ensure the integrity and accuracy of air pollution information.

Acknowledgements The authors thank ARTA for providing the data. This research was funded by project PID2020-113961GB-I00 of the Spanish Ministry of Science and Innovation (also supported by the FEDER program), project FQM-307 of the Government of Andalusia (Spain) and the PhD grant (FPU18/01779) awarded to Christian Acal. The authors also thank the support of the University of Granada, Spain, under project for young researchers PPJIB2020-01.

References

1. Wang, P., Chen, K., Zhu, S., Wang, P., Zhang, H.: Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak. Resour. Conserv. Recy. **158**, 104814 (2020) doi: 10.1016/j.resconrec.2020.104814
2. Berman, J. D., Ebisu, K.: Changes in U.S. air pollution during the COVID-19 pandemic. Sci. Total. Environ. **739**, 139864 (2020) doi:https://doi.org/10.1016/j.scitotenv.2020.139864
3. Sicard, P., De Marco, A., Agathokleous, E., Feng, Z., Xu, X., Paoletti, E., Rodriguez, J. J. D., Calatayud, V.: Amplified ozone pollution in cities during the COVID-19 lockdown. Sci. Total. Environ. **735**, 139542 (2020) doi: https://doi.org/10.1016/j.scitotenv.2020.139542
4. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag, New York (2002)
5. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, second ed. Springer-Verlag, New York (2005)
6. Escabias, M., Aguilera, A. M., Valderrama, M. J.: Modeling environmental data by functional principal component logistic regression. Environmetrics **16**, 95-107 (2005) doi: https://doi.org/10.1002/env.696
7. Martínez Torres, J., Pastor Peréz, J., Sancho Val, J., McNabola, A., Martínez Comesaña, M., Gallagher, J.: A Functional Data Analysis Approach for the Detection of Air Pollution Episodes and Outliers: A Case Study in Dublin, Ireland. Mathematics **8**, (2020) doi:/doi.org/10.3390/math8020225
8. Smaga, L.: A note on repeated measures analysis for functional data. AStA Adv. Stat. Anal. **104**, 117-139 (2020) doi: https://doi.org/10.1007/s10182-018-00348-8
9. Acal,C., Aguilera, A. M., Sarra, A., Evangelista, A., Di Battista, T., Palermi, S.: Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. Stoch. Environ. Res. Risk. Assess. 1-19 (2021) doi: https://doi.org/10.1007/s00477-021-02071-4
10. Ocaña, F. A., Aguilera, A. M., Escabias, M.: Computational considerations in functional principal component analysis. Comput. Stat. **22**, 449-465 (2007) doi:10.1007/s00180-007-0051-2
11. Aguilera, A. M., Acal, C., Aguilera-Morillo, M. C., Jimènez-Molinos, F., Roldàn, J. B.: Homogeneity problem for basis expansion of functional data with applications to resistive memories. Math. Comput. Simul. **186**, 41-54 (2021) doi: https://doi.org/10.1016/j.matcom.2020.05.018

Conformal Prediction for Geographically Weighted Functional Regression models: an application for environmental impact assessment.

Conformal Prediction per modelli di regressione geografica funzionale: un'applicazione per la valutazione dell'impatto ambientale

Andrea Diana, Elvira Romano and Antonio Irpino

Abstract In this work we introduce a general framework for distribution-free predictive inference in Geographically Weighted Functional Regression (GWFR) using conformal inference. A prediction band for the functional response variable using a functional estimator of the regression function and a new non conformity measure is thus introduced. We also investigate our procedure for producing prediction bands with locally varying length, in order to adapt the approach in presence of heteroskedasticity in the data. An application for environmental impact assessment is proposed.

Abstract In questo lavoro viene introdotto un metodo di inferenza non parametrica per la validazione di un modello di Regressione Geografica Funzionale. In particolare viene proposta una banda di predizione per la variabile di risposta funzionale utilizzando uno stimatore funzionale ed una nuova misura di conformitá. Le caratteristiche del metodo proposto vengo illustrate mediante un'applicazione per la valutazione d'impatto ambientale.

Key words: functional data, spatial dependence, conformal prediction, geographical weighted regression

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

Antonio Irpino
University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: antonio.irpino@unicampania.it

1 Introduction

In the last decade Functional Data Analysis (FDA) [7], [9] has been the focus of much research efforts in the statistics and machine learning community. The core of FDA consists in considering functions rather than scalars or vectors as object of the analysis. This approach provides a powerful modelling tool for many and many natural processes with certain smoothness structures, typical examples arise from medicine, biomedicine, public health, biology, biomechanics and environmental science, etc. Literature on functional data analysis is growing very quickly and familiar techniques like regression have been extended to functional data. A crucial challenge for these models is quantifying uncertainty in prediction. Methods related to this framework consists in working on parametric bootstrapping techniques ([2], [4]), or in the application of dimensional reduction techniques to manage the naturally infinite dimensional problem ([1], [8]). More recent works are based on a novel approach to forecast by using Conformal Prediction (CP) ([6]). This last one overcome the previous literature since it is able to output either exact or valid prediction bands under minimal distributional assumptions.
The objective of the present work is to introduce a general framework for distribution-free predictive inference in Geographically Weighted Functional Regression (GWFR) using conformal inference. A prediction band for the functional response variable is proposed and a new non conformity measure is introduced. Prediction bands with locally varying length are defined, in order to adapt the approach in presence of heteroskedasticity in the data. An application for environmental impact assessment is discussed.

2 GWR for geostatistical functional data

Geostatistical functional data $(X_{s_1}(t),\ldots,X_{s_i}(t),\ldots,X_{s_n}(t))$ are random functions $X_s(t)$ located in n points $(s_1,\ldots,s_i,\ldots,s_n)$ in $D \subseteq R^d$. Each function is defined on $T=[a,b] \subseteq R$ and is assumed to belong to a Hilbert space with the inner product $\langle X_{s_i}, X_{s_j}\rangle = \int_T X_{s_i}(t)X_{s_j}(t)dt$ [9]. For a fixed site s_i, it is assumed that the observed functions can be expressed according to the model: $X_{s_i}(t) = \mu_{s_i}(t) + \varepsilon_{s_i}(t), \quad i = 1,\ldots,n$ where $\varepsilon_{s_i}(t)$ are zero-mean residuals and $\mu_{s_i}(t)$ is the mean function.
For each $t, t \in T$, the random process is assumed to be second order stationary and isotropic: that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites. It is assumed that the mean function is constant over D and that the semivariogram function $\gamma(h,t) = \gamma_{s_i s_j}(t) = \frac{1}{2}V(X_{s_i}(t) - X_{s_j}(t))$, according to [5], can be expressed by:

$$\gamma(h,t) = \gamma_{s_i s_j}(t) = \frac{1}{2}V(X_{s_i}(t) - X_{s_j}(t)) = \frac{1}{2}E\left[X_{s_i}(t) - X_{s_j}(t)\right]^2. \qquad (1)$$

Let's suppose we want to predict a functional response variable $Y_s = \{Y_s(\tau), \tau \in T_1\}$ starting by K functional covariates $\chi_{s_i}(t) = [X_{s_{i,1}}(t), \ldots, X_{s_{i,K}}(t)]^T$. We can consider the geographically weighted regression (GWR) model [11] given by

$$Y_{s_i}(\tau) = \beta_0(\tau, s_j) + \sum_{k=1}^{K} \int_T X_{s_i,k}(t)\beta_k(t, \tau, s_j)dt + \varepsilon_{s_i}(\tau),\ i = 1, \ldots, n, \qquad (2)$$

where the function $\beta_0(\tau, s_j)$ is the mean function at location s_j, $\beta_k(t, \tau, s_j)$ is the regression function for the $k-$th covariate at location s_j, and $\varepsilon_{s_i}(\tau)$ is a random error function at point s_i. Statistical inference on this model is strictly related on testing the variability of the coefficients and in constructing confidence interval by using bootstrapping procedures ([11]). In the following we introduce a conformal approach to construct prediction bands such to have finite sample validity, without assumption on the probability models.

3 Conformal Prediction for GWR

Consider i.i.d. regression data $(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau)) \sim P$, where each $(\chi_{s_i}(t), Y_{s_i}(\tau))$ is a multivariate spatial functional stochastic process or a multivariate functional random field in $\mathscr{L}_2(T)^K \times \mathscr{L}_2(T_1)$, comprised of a response variable $Y_{s_i}(\tau)$ and a k$-$dimensional vector of features (or predictors, or covariates) $\chi_{s_i}(t) = (X_{s_i,1}(t), \ldots, X_{s_i,K}(t))$. The feature dimension K may be large relative to the sample size n (in an asymptotic model, K is allowed to increase with n). We are interested in predicting a new response $Y_{s_{n+1}}(\tau)$ from a new feature value $\chi_{s_{n+1}}(t)$, with no assumptions on P. Formally, given a nominal miscoverage level $\alpha \in (0,\ 1)$, a prediction band $C \subset \mathscr{L}_2(T)^k \times \mathscr{L}_2(T_1)$ based on $(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau))$ is such that

$$\mathbb{P}(Y_{s_{n+1}}(\tau) \in C(\chi_{s_{n+1}}(t))) \geq 1 - \alpha, \qquad (3)$$

where the probability is taken over the s_{n+1} i.i.d. draws $(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau), (\chi_{s_{n+1}}(t), Y_{s_{n+1}}(\tau))) \sim P$, and $C(\chi_s(t)) = \{Y_s(\tau) \in \mathscr{L}_2(T) :\ (\chi_s(t), Y_s(\tau)) \in C\}$ for a point $\chi_s(t) \in \mathscr{L}_2(T)^k$. The defined prediction bands have finite-sample (nonasymptotic) validity, without assumptions on P. These can be obtaining by defining the following algorithm of conformal prediction:

Algorithm of conformal prediction

Input: Data $(\chi_{s_i}(t), Y_{s_i}(\tau)),\ i = 1, \ldots, n$, miscoverage level $\alpha \in (0,1)$, Nonconformity measure $\mathscr{D}$, regression algorithm $\mathscr{A}$, points $\chi_{new}(t) = \{\chi_{s_{n+1}}(t), \chi_{s_{n+2}}(t), \ldots\}$ at which to construct prediction band, and values $Y_{trial}(\tau) = \{Y_{s_{n+2}}(\tau), Y_{s_{n+2}}(\tau), \ldots\}$ to act as trial values
Output: Predictions band, at each element of $\chi_{new}(t)$
for $\chi_s(t) \in \chi_{new}(t)$ **do**

for $Y_s(\tau) \in Y_{trial}(\tau)$ **do**
$\hat{Y}_s(\tau) = \mathscr{A}(\{(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau)), (\chi_s(t), Y_s(\tau))\})$
$R_{Y_s(\tau),s_i} = \mathscr{D}(\hat{Y}_s(\tau), Y_{s_i}(\tau)), i = 1, \ldots, n$ and $R_{Ys,s} = \mathscr{D}(\hat{Y}_s(\tau), Y_s(\tau))$
$\pi(Y_s(\tau)) = \frac{1+\sum_{i=1}^{n} \mathbf{1}\{R_{Y_s(\tau),s_i} \leq R_{Y_s(\tau),s}\}}{n+1}$
end for
$C_{conf}(\chi_s(t)) = \{Y_s(\tau) \in Y_{trial}(\tau) : (n+1)\pi(Y_s(\tau)) \leq (1-\alpha)(n+1)\}$
end for
Return $C_{conf}(\chi_s(t)), for\ each\ \chi_s(t) \in \chi_{new}(t)$

The regression algorithm $\mathscr{A}$ in our case is GWR for geostatistical functional data. The non-conformity measure $\mathscr{D}$ is an optimally weighted distance for functional data spatially dependent defined as in [10]

$$d_{\omega_s}\left(X_{s_i}(t), X_{s_j}(t)\right) = \sqrt{\int_T \omega_s(t)(X_{s_i}(t) - X_{s_j}(t))^2 dt} \tag{4}$$

where the weight ω_s satisfies $\omega_s \geq 0$ and $\int \omega_s dt = 1$. Using distance in 4, we consider weight functions including both the spatial and functional component. It is a generalisation of [3] to the spatial functional framework for two different spatial domains: the georeferenced and the directed network.
The spatio-functional smooth function is obtained by the following minimisation problem:

$$\omega_s(t) = argmin_{||\omega_s||=1} \frac{\sum_{1 \leq i < j \leq n} V(||\theta_{i,j})||^2_{\omega_s})}{\sum_{1 \leq i < j \leq n} [E(||\theta_{i,j})||^2_{\omega_s})]^2}; \tag{5}$$

with $\theta_{i,j}(t) = a_{i,j}X_i(t) - a_{j,i}X_j(t)$, where $a_{i,j}$ and $a_{j,i}$ are obtained starting from the structure of the spatial domain of interest. The coefficient $a_{j,i}$ is the element reflecting the spatial dependence among functional data and changes according to the spatial grid on which the functional data are observed. In our case, we choose a weight function depending on the spatial variability expressed by a trace-variogram function. Formally we define: $a_{i,j} = a_{j,i} = \hat{\gamma}(h_{i,j})$ where $\hat{\gamma}(h)$ is the estimated trace-variogramm.

4 Air quality modelling via GWR

In environmental studies, one of the main challenges is to monitor the effect of air pollution on the health by measuring daily exposure to the pollutant. Particulate matter (PM) in different concentrations and Sulfur dioxide (SO_2) are the classical used variables to evaluate and estimate health effect of exposure. Studies in this framework usually consider the monitored ozone concentration to a daily sum-

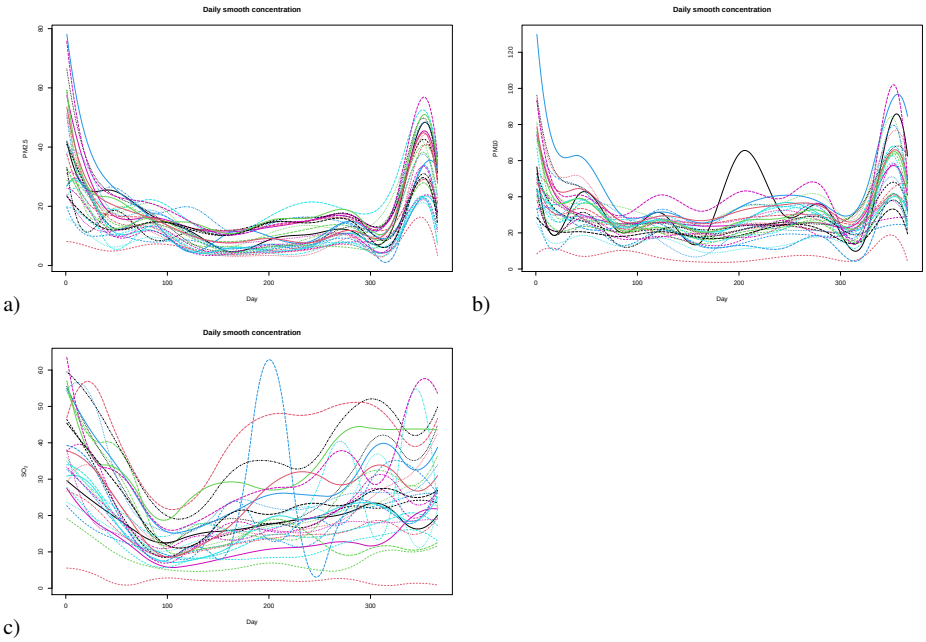


Fig. 1 Daily smooth concentration curves of $PM2.5$ (a), $PM10$ (b) and SO_2 (c), respectively.

mary measure and estimate health effect by regressing day-varying health outcome against day-varying summary measure of ozone.

In this work we suggest to use the functional form of ozone concentration and propose to estimate the effect of daily exposure by using a GWR model. We assume that a daily smooth concentration curve represents a measure of daily exposure to the concentration and adopt georeferenced functional regression method to estimate the effect of the exposure using the concentration curves of $PM10$ and SO_2 as predictors and $PM2.5$ as response variable. We thus explore how the total daily concentration of $PM2.5$ depends on the specific features of the ozone profile of a day (Fig.1).

For the application, we use data from Campania Region (Italy) over the year 2020 (www.arpacampania.it).The analysis was performed using the free available R software.

Since the aim is to construct valid, meaningful and interpretable prediction bands we started by these two covariates, such to identify by the periods of time in which the high simultaneous variability of $PM10$ and SO_2 could become critical in a the day of the week and can led to modify the impact on the $PM2.5$.

At first we fitted the model specified in (2) using the $PM10$ and SO_2 measure as a predictors and the $PM2.5$ as functional response.

Then a conformal prediction step has been performed. The results show the dependence of the concentration of $PM2.5$ on exposure to $PM10$ and SO_2. This dependence is not significative as can be seen in the conformal prediction step. Starting

Table 1 The table shows the percentage (average on 100 simulations) of the curves $\hat{Y}_s(\tau)$ $(PM2.5)$ that are in their prediction band.

Simple Length	Trial Length	α	Percetnege
20	10	0.05	1.6%
20	10	0.1	8%
25	5	0.05	3%
25	5	0.1	9.3%

by varying the size of the sample and the trial length, 100 simulation show how the percentage of error in the conformal prediction bands is high respect to the fixed miscovarage level α, as can be seen in Table1.

This suggests the introduction of other possible carefully chosen covariates. Then the model is thus been investigated by considering further climatic variables. Results seem encouraging and show how the proposed conformal prediction step represents a fundamental step in a predictive framework.

References

1. Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J.M.: A prediction interval for a function-valued forecast model: Application to load forecasting. Int. J. Forecast. **32**, 939–947, (2016)
2. Cao, G., Yang, L., Todem, D.: Simultaneous Inference For The Mean Function Based on Dense Functional Data. J. Nonparametr. Stat. **24**, 359–377 (2012)
3. Chen H., Reiss, P.T., Tarpey, T.:Optimally Weighted L2 Distance for Functional Data. Biometrics **70(3)**, 516–525, (2014)
4. Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Statist. Sinica **21**, (2011)
5. Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric **21**, 224–239, (2010)
6. Diquigiovanni, J., Fontana, M., Vantini, S.: The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. arXiv:2102.06746 (2021)
7. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Verlag (2006)
8. Hyndman, R.J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: A functional data approach. Comput. Statist. Data Anal. **51**, 4942–4956, (2007)
9. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2005)
10. Romano, E., Diana, A., Miller, C., ODonnell, R.: Optimally weighted L2 distances for spatially dependent functional data. Spatial Statistics **39**, (2020)
11. Yamanishi, Y. and Tanaka, Y.: Geographically weighted functional multiple regression analysis: A numerical investigation. Journal of Japanese Society of Computational Statistics **15**, 307–317, (2003).

Session of solicited contributes SS21 – *Local sustainability assessment: challenges in building quality indicators*
Organizers and Chairs: Francesca Fortuna & Alessia Naccarato

Urban Sustainability Assessment: A Proposal for an Index Based on SDGs' Indicators

La misurazione della sostenibilità urbana: una proposta per un indice composito basato sugli indicatori dei Sustainable Development Goals

Elena Grimaccia

Abstract Urban sustainability is at the centre of the global debate on development challenges. Measuring progress towards sustainable urban development requires specific quantification, and numerous initiatives have been developed for monitoring and comparing the sustainability performance of urban areas. In order to assess urban sustainability, this paper employs the Indicators identified among the Goal 11 of the Sustainable Development Goals, and analyse the latent dimension underlying the construct of "urban sustainability" as foreseen by UN, through indices based on different latent variable models. The paper demonstrates that the indicators chosen by the UN to monitor urban sustainability are indeed related to two different latent constructs: urban sustainability and impact of natural disaster.

Abstract *La sostenibilità urbana è al centro del dibattito globale sulle sfide dello sviluppo. La misurazione dei progressi verso uno sviluppo urbano sostenibile richiede una quantificazione specifica, e sono state sviluppate numerose iniziative per monitorare e confrontare le aree urbane in termini di sostenibilità. Per valutare la sostenibilità urbana, il presente lavoro utilizza gli Indicatori del Goal 11 degli Obiettivi di Sviluppo Sostenibile e analizza la dimensione latente alla base del costrutto di "sostenibilità urbana" previsto dall'ONU, attraverso indici basati su diversi modelli di variabili latenti. Il lavoro dimostra che gli indicatori scelti dall'ONU per monitorare la sostenibilità urbana sono legati a due diversi costrutti latenti: sostenibilità urbana e impatto dei disastri naturali.*

Key words: Structural Equation Models (SEM), Multiple Indicator Multiple Causes (MIMIC) models, Urban Sustainability, Cities, Sustainable Development Goals (SDGs).

[1] Elena Grimaccia, Istat; email: elgrimac@istat.it

1 Introduction

The role of cities in achieving sustainability is of key importance and today it is at the centre of the scientific and public debate, since more than a half of world population lives in cities or urban settlements (Sharifi, 2020), cities will be home to 60% of the global population, and this share is estimated to further increase to about 68% by 2050 (UNDESA, 2018). This means that, from 2010 to 2050, between 2.5 to 3 billion people will be added to the urban population worldwide. Nowadays, cities already contribute about 80% of global GDP. However, cities also account for about 70% of global energy consumption, 70% of global carbon emissions, as well as over 70% of resource use, making it even more compelling for policy makers to design sustainable policies for urban settlements (UNHABITAT, 2020).

The assessment of urban sustainability provides a tool for policy makers to monitor progress towards sustainable development of urban areas, improving transparency of decision making, and increasing awareness (Sharifi, 2019). Indeed, the United Nations (UN) chose to dedicate an entire Goal of the Agenda 2030 to urban sustainability: the Sustainable Development Goal 11 is known as the 'urban SDG: to make cities and human settlements inclusive, safe, resilient and sustainable' (UN, 2015). The SDG 11 consists of a number of targets, including: adequate, safe and affordable housing; accessible and sustainable transport systems for all; inclusive and sustainable urbanisation; to reduce the number of people affected by disasters; to reduce the environmental impact of cities (UN, 2017).

As a multidimensional phenomenon, Urban Sustainability needs to be assessed into a simple and readable measure. Among the huge range of measurement proposals that have been developed for monitoring and comparing urban sustainability (Merino-Saum et al., 2020), the paper refers to UN experience, since the international process to identify the theoretical framework at the basis of the SDG indicators benefit from a wide accepted approach. A Composite indicator (CI) of Urban Sustainability is built with the aim of capturing different and relevant aspects of a latent multidimensional reality (Becker et al., 2017). CIs are able to summarise multidimensional phenomena, so that they are useful tools to ease interpretation and to allow benchmarking. However, care must be taken in their construction and interpretation to avoid misleading policy messages (Floridi et al., 2011).

The present study was designed to extend previous research by comparing results from different latent variable models, and employing in particular Multiple Indicator Multiple Causes (MIMIC) models (Joreskorg and Goldberger, 1975; Bollen, 1989; Krishnakumar and Nagar, 2008), which allow for simultaneous evaluation of correlations between the latent construct(s) and among the indicators and the explanatory variables. A Composite Indicator is obtained, based on the estimate of the latent concepts, modelling causal relationships among observable indicators (Krishnakumar and Nagar, 2008; Lauro et al., 2018).

2 Data and Methods

The present paper proposes an Urban Sustainable Development Index, based on the aggregation of the indicators included in the SDG11, applying a Principal Component Analysis, and subsequently estimating different MIMIC models.

The availability of data referring to the SDGs targets' Indicators is still an issue. However, 162 countries have been considered in the analysis.

The selected target indicators assess different aspects of urban sustainability: the proportion of urban population living in slums, informal settlements or inadequate housing ("Slums", negative polarity); the proportion of urban solid waste regularly collected ("Waste"); the annual mean levels of fine particulate matter (e.g. PM2.5 and PM10) in cities, population weighted ("Particulate", negative polarity); the number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population ("Disasters", negative polarity); direct economic loss in relation to global GDP, damage to critical infrastructure and number of disruptions to basic services, attributed to disasters ("GDP-Disaster", negative polarity); and, finally, the proportion of local governments that adopt and implement local disaster risk reduction strategies, in relation with the Sendai Framework for Disaster Risk Reduction ("Sendai").

The measurement model is, in this case, formative, since the latent construct is formed by a combination of indicators: given the nature and the correlations of our manifest variables (Grimaccia et al., 2020), and being the available indicators not highly correlated and also scarcely interchangeable, a formative model is more suitable (Booysen, 2002; Hoyle, 2012).

Very briefly, MIMIC models (Joreskorg and Goldberger, 1975), modelling relationships among observed and unobserved (latent) variables, take into account two levels of relations: the first one is the measurement model in which latent variable is measured by means of a set of observable variables; the second level considers the causal relations among the latent variable(s) and auxiliary variables. MIMIC model is founded upon the specification of a system of equations which specify the relationship between an unobservable latent variable (urban sustainability), a set of observable endogenous indicators and a set of observable exogenous variables, that are believed to be the causes of urban sustainability.

3 Results

An Exploratory Principal Component Analysis (Krishnakumar and Nadar, 2008) is conducted to determine the manifest variables that measure the latent factors for urban

sustainability. The analysis allows to represent the relationships among the (manifest) variables. The results show the presence of two latent factors (Figure 1.A).

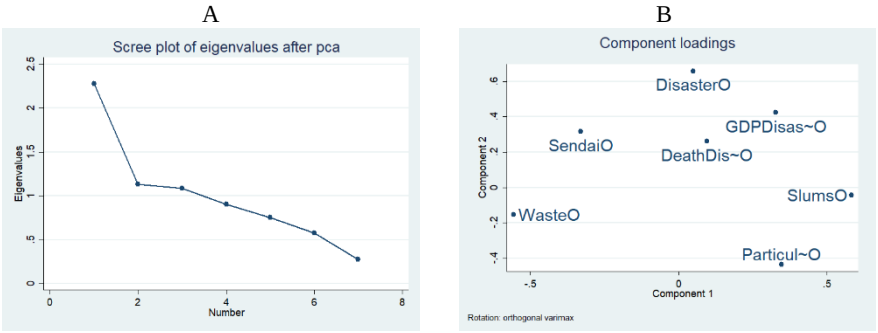


Figure 1: PCA eigenvalues and factor loadings.

The loadings in Figure 1.B indicate that the first latent construct largely determines the Urban Sustainability concept, and includes the variables on the management of Waste on the negative side of the axis, and the level of pollution due to particulate and the population living in slums on the positive side of the horizontal axis. The second latent construct corresponds to the Disaster component and has a lower impact on the overall variability.

A MIMIC model was used to determine the degree of association between the manifest variables and our latent concept.

In this model of Urban Sustainability, the formative construct is composed by the identified indicators on Particulate matter concentration, Waste recycling rate, Slums and Disasters, Sendai Policies, and GDP lost in disaster, while the exogenous variables (or causes) of the model are identified by the three indicators forming the United Nation's Human Development Index (HDI). The HDI summarizes economic and social aspects of the level of development (Anand and Sen 1997), and it is an established measure of human development that is related to urban sustainability analysis (Merino-Saum et al., 2020). Instead of using the thresholds defined for the HDI, that presents some subjective choices in the aggregation process, the three indicators separately have been considered, having verified that the correlations among them are very high (Figure 2). As underlined in Merino-Saum et al. (2020), Urban Sustainability is related to Life expectancy, Income (per capita) and Level of education, measured by "Mean years of schooling".

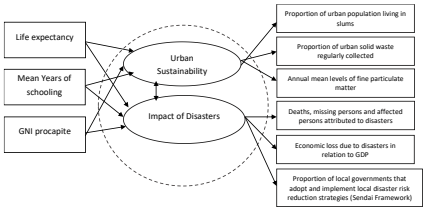


Figure 2: Methodological Framework for Urban Sustainability Assessment.

Since the formative construct needs to be theoretically and conceptually refined by exchanging and or adding indicators (Hair et al., 2017), different models with various specifications have been estimated.

A first MIMIC model with a single latent factor has been estimated. In relation to the latent factor, which could be defined as "Urban sustainability", the presence of slums and pollution have a high negative impact, while the management of waste present a positive weight. The indicators related to disasters have a lower impact on the latent factor, and the coefficient referring to the Sendai framework is not significant. These results are confirmed by the estimation of the same model, excluding the variable "Sendai".

These results are coherent with the preliminary analysis conducted by a PCA, that provided indication on the existence of two different latent factors: the actual urban sustainability and a construct related to the effects of natural disaster.

Including the GNI per capita as a "mediator" variable in the model, together with the four identified urban sustainability variables, the estimates present very good levels of the indexes of goodness of fit, and interesting results. The levels of people in slums, of particulate in the air, and the number of victims in disasters all present a significant impact on the Urban Sustainability construct, and the GNI per capita is significantly related to the latent construct. Including in the model the variable regarding the level of education (Mean years of schooling), does not improve the overall quality of the model, and the variable is not even significant.

Very interesting are, instead, the results of a model where life expectancy is included in the formative model. In this latter specification, Slums, Waste, and Particulate all show significant impact on the latent construct of "Urban Sustainability", together with Life expectancy, with the GNI as a (significant) mediator variable.

5. Conclusions

The paper demonstrates that the indicators chosen by the UN to monitor the SDG on urban sustainability are indeed linked to two different latent constructs: urban sustainability and impact of natural disaster. This creates a lot of difficulties in building a unique composite indicator for the two dimensions. The above mentioned latent constructs are however related and both are linked to the level of development of the country measured by the indicators composing the HDI. The choice made by the stakeholders of the UN of including the effects of natural disaster is arguable also because the indicators related to the impact of disasters is taken into account (with the same indicators, e.g. deaths, GDP lost, and so on) in other two SDGs: Goal 1 ("End poverty") and in Goal 8 on sustainable and equitable growth.

However, both applying the weights of a FCA and those based on a MIMIC specification, composite indicators of Urban Sustainability have been obtained. Such indexes would be useful to establish a ranking among Nations towards Urban Sustainability. Further research should be devoted to compare different additional specification of the model and the related ranking of countries.

References

1. Anand, S., Sen, A.K.: Concepts of human development and poverty: A multidimensional perspective. Human Development Papers, United Nations Development Programme, New York (1997)
2. Becker, W., Saisana, M., Paruolo, P., Vandecasteele, I.: Weights and importance in composite indicators: Closing the gap. Ecological Indicators **80** (2017)
3. Bollen, K. A.: Structural equations with latent variables. John Wiley & Sons (1989).
4. Booysen, F.: An Overview and Evaluation of Composite Indices of Development. Social Indicators **59**, 115–151 (2002)
5. Floridi, M., Pagni, S., Falorni, S., Luzzati, T.: An exercise in composite indicators construction: Assessing the sustainability of Italian regions. Ecological Economics **70**, 1440–1447 (2011)
6. Grimaccia, E., Naccarato, A., Terzi, S.: World ranking of urban sustainability through composite indicators. In: Pollice, A., Salvati, N., Schirripa Spagnolo, F. (eds.) Book of Short Papers SIS 2020. pp. 1017-1022, Pearson (2020).
7. Hair, J., Hollingsworth, C. L., Randolph, A. B., Chong, A. Y. L.: An updated and expanded assessment of PLS-SEM in information systems research. Industrial Management & Data Systems **117** (3) (2017)
8. Hoyle, R.H.: Handbook of structural equation modeling. Guilford Press (2012)
9. Joreskorg, K. G., Goldberger, A. S.: Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. Journal of the American Statistical Association **70** (351), 631-639 (1975)
10. Krishnakumar, J., Nagar, A. L.: On Exact Statistical Properties of Multidimensional Indices Based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models. Social Indicators Research **86**, 481–496 (2008)
11. Lauro, C. N., Grassia, M. G., Cataldo, R.: Model based composite indicators: New developments in partial least squares-path modeling for the building of different types of composite indicators. Social Indicators Research **135**(2), 421–455 (2018)
12. Merino-Saum, A., Halla, P., Superti, V., Boesch, A., Binder, C.R.: Indicators for urban sustainability: Key lessons from a systematic analysis of 67 measurement initiatives. Ecological Indicators **119**, 106879 (2020)
13. Sharifi, A.: A critical review of selected smart city assessment tools and indicator sets. Journal of Cleaner Production **233**, 1269-1283 (2019)
14. Sharifi, A.: Urban sustainability assessment: An overview and bibliometric analysis. Ecological Indicators **121**, 107102 (2020)
15. UN General Assembly: Resolution adopted by the General Assembly on 25 September 2015 n. 70/1. Transforming our world: the 2030 Agenda for Sustainable Development. New York (2015)
16. UNHABITAT: World Cities Report 2020. The Value of Sustainable Urbanization (2020)
17. United Nations Statistical Commission: Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. UN Resolution A/ RES/71/313 (2017)
18. United Nations, Department of Economic and Social Affairs, Population Division: World Urbanization Prospects: The 2018 Revision, Online Edition. File 21: Annual Percentage of Population at Mid-Year Residing in Urban Areas by Region, Sub-region, Country and Area, 1950-2050 (2018)

Between and Within Country Inequality in Regional Well Being

Disuguaglianze Between e Within nel benessere regionale dei paesi dell'OCSE

Paolo Liberati and Giuliano Resce

Abstract This paper analyses the inequality between the regions of the OECD by the stochastic multi-objective acceptability analysis and the associated multivariate Gini index. The distribution of the potential rankings for each region is used to measure multidimensional inequality both within and between countries. Beyond the expected two clubs of rich and poor countries, a third group of countries emerges that belongs neither to the top nor to the bottom of ranking, with significant economic differences among regions. Overall, the bulk of inequality is between countries, and we find an inverse U-shape connection between the regional well-being and its inequality within the OECD member countries.

Abstract *Questo articolo analizza la disuguaglianza tra le regioni dei paesi membri dell'OCSE mediante un approccio multi-criteriale e l'indice di Gini multivariato. Per misurare la disuguaglianza multidimensionale sia all'interno che tra i paesi si usa la distribuzione delle classifiche potenziali per ciascuna regione. Oltre ai ben noti due club di paesi ricchi e paesi poveri, emerge un terzo gruppo di paesi che non appartiene né alla parte superiore né alla parte inferiore della classifica, con significative differenze economiche tra le regioni. In generale, la maggior parte della disuguaglianza è tra i paesi e troviamo una connessione prima crescente poi decrescente tra il benessere regionale e la sua disuguaglianza all'interno dei paesi membri dell'OCSE.*

Key words: Regional Well-being; Multidimensional Inequality; OECD

[1] Paolo Liberati, Roma Tre University; e-mail: paolo.liberati@uniroma3.it: Giuliano Resce, University of Molise; e-mail: giuliano.resce@unimol.it

Footnotes

1. Introduction

Among the alternative measures of well-being which have been proposed by international institutions, as well as by national statistical offices (Costanza et al. 2014; 2016), the Better Life Initiative (BLI) launched by the Organisation for Economic Co-operation and Development (OECD) in 2011, is one of the most popular (Durand, 2015; Decancq, 2017).

The BLI is based on the recommendations made in Stiglitz et al. (2010), that well-being is multidimensional and has different key aspects of life that should be considered simultaneously. As many of the policies that affect people's lives are local or regional, in 2014 OECD launched the "Regional Well-Being" project (OECD, 2014), where each of the 402 regions of the 36 countries included in BLI is measured in eleven topics – income, jobs, housing, health, access to services, environment, education, safety, civic engagement and governance, community, and life satisfaction. Taking advantage of this detailed dataset, this paper proposes to analyse the differences in the regional well-being using the recent developments both in the multidimensional inequality measures and in the criteria to overcome the use of a specific set of weights to aggregate multidimensionality in a single indicator.

Previous literature on regional well-being is affected by the shortcoming represented by the choice of a specific vector of weights. (Peiró-Palomino, 2019; Peiró-Palomino et al. 2019; Döpke et al. 2017), and almost nothing is known about whether ranking would change when changing the vector of weights, which is instead a fundamental information for the analysis of multidimensional well-being. Our analysis explicitly deals with this issue and contributes to the literature by proposing the use of the Stochastic Multi-Objective Acceptability Analysis (SMAA) approach and its inequality metrics recently proposed in Greco et al. (2018), Lagravinese et al. (2019), and Coco et al. (2020). This technique allows considering a large set of vectors of weights to derive alternative feasible ranks of the regions. SMAA determines, for each region, the probability to have a given rank in the overall ranking, a measure that will be referred to as its rank acceptability index. These ranks are finally used to decompose multidimensional well-being within and between countries (Lagravinese et al. 2019). To this purpose, we replicate the calculus of the regional multidimensional well-being index by using 10,000 different vectors of weights, where all regions, in each replication, are ranked using a common set of weights. Finally, the decomposition of well-being is carried out by first using a generalisation of the Gini index applied to the rank acceptability index, and then decomposing it using the ANalysis Of GIni (ANOGI) decomposition in order to get further information also on the degree of overlapping between the distributions of well-being. Previous SMAA applications in economics have been focussed on regional well-being within a single country (Greco et al. 2018), the comparisons in sustainable development between European Countries (Resce and Fritz 2021), and the assessment of education performances in OECD member countries (Coco et al. 2020). By leveraging the regional well-being data (OECD, 2014) and the recent advances in ANOGI (Lagravinese et al. 2019), this paper for the first time uses SMAA for comparing regional well-being across countries.

2. Data and Methods

Regional measures are available for OECD regions in eleven well-being indicators: income, jobs, housing, education, health, environment, safety, civic engagement and governance, access to services, community, and life satisfaction. For each topic, one or two indicators have been selected by OECD (2014). Since well-being indicators are expressed in different units, to compare indicators on the same scale, they have been normalised using the min-max method (Nardo et al. 2008), to have values ranging from 0 to 10. To reduce the skewness of the distribution, a threshold has been applied by OECD (2014) to eliminate those values that are below the 4th percentile and above the 96th percentile. In the case of the homicide rate, since only few regions have a very high value, the cut-offs are the 10th and the 90th percentile, respectively. Imposing a threshold on extreme values allows to obtain well-being scores that are more evenly distributed and avoids cases where almost all regions would be included between values 9 and 10 (OECD, 2014). For this analysis, we have used the last version of OECD Regional Well-Being indicator data file[1]. For analytical purposes, the OECD classifies regions as the first administrative tier of sub-national government. This classification is used by the National Statistical Offices to collect information and it represents the framework for implementing regional policies in many countries. While the number of regions varies from country to country, the international comparability is ensured by the fact that these administrative regions are officially established in countries (OECD, 2014).
In the regional well-being measure proposed by OECD (2014), a set of m (402) regions $A = \{a_1, \dots, a_m\}$ is evaluated on a set of n (11) topics $G = \{g_1, \dots, g_n\}$. The most common way to deal with multidimensionality is to aggregate the set of elementary indices either using a simple unweighted arithmetic mean or by assigning a specific set of different weights to the elementary indices. In what follows, for each region $a_k \in A$, an overall evaluation function $u(a_k, w)$ depending on the chosen vector of weights $w = w_1, \dots, w_n$ can be defined as follows:

$$u(a_k, w) = \sum_{i=1}^{n} w_i g_i(a_k) \qquad (1)$$

Clearly, the ranking of m regions is dependent on the chosen weights $w_1, \dots, w_n$. The Stochastic Multi-Objective Acceptability Analysis (SMAA) (Lahdelma *et al.*, 1998; Lahdelma and Salminen, 2001) takes this point explicitly into account, as it allows calculating the probability that region a_k has the r-th position in the ranking (b_k^r). Using b_k^r we quantify the probability of having a rank equal to or higher than a specific threshold (80%) as the Upward Cumulative Rank Acceptability Index (UCRAI), and the probability of having a rank equal to or lower than specific threshold (20%) as the Downward Cumulative Rank Acceptability Index (DCRAI). UCRAI and DCRAI represent, in respectively, the probability that each region is among the worst and among the best performers.

[1] https://www.oecdregionalwellbeing.org/

3. Results

Our analysis shows that different rankings may be obtained according to the vector used to weight the eleven metrics. Some regions can be classified either among those with the highest level of well-being or among those with the lowest level of well-being, depending on the specific vector of weights. This outcome occurs most likely in those countries where there is large heterogeneity of the eleven metrics among regions (Canada, Cech Republic, Italy, Japan, Korea, New Zealand, Spain, and United States), something that is connected to the internal dualism among different territories of the same country (Table 1). In countries where this internal dualism in negligible – and thus the internal heterogeneity among territories is low – some regions usually belong either to the top or to the bottom of ranking. The previous literature has shown that the internal dualism is among the most important determinants of country differences in income inequality, and it has a negative association with the performance measured by income (Iammarino et al., 2019; Bourguignon, Morrisson, 1998; Alesina and Rodrik, 1994; Krugman, 1991). Our results contribute to these findings showing that the negative association between regional inequality and performance is confirmed in the multidimensional context. Furthermore, our evidence reveals an inverse U-shape connection, a kind of cross-sectional Kuznetz curve (Kuznets, 1955) between the regional well-being and its inequality in the OECD member countries. From a policy perspective, these results suggest that a low degree of inequality is not inconsistent with a high level of well-being.

Table 1: Descriptive statistics of DCRAI (probabilities of being among the top 100 regions) and UCRAI (probabilities of being among the bottom 100 regions) by country

Country	N	DCRAI			UCRAI		
		mean	G	O	mean	G	O
Australia	8	0.997	0.003	0.132	0.997	0.003	0.132
Austria	9	0.595	0.27	0.626	0.595	0.27	0.626
Belgium	3	0.313	0.664	0.722	0.313	0.664	0.722
Canada	13	0.591	0.342	0.772	0.591	0.342	0.772
Chile	15	0	.	.	0	.	.
Czech R.	8	0	0.875	0.619	0	0.875	0.619
Denmark	5	0.811	0.138	0.308	0.811	0.138	0.308
Estonia	5	0.001	0.8	0.624	0.001	0.8	0.624
Finland	5	0.501	0.332	0.32	0.501	0.332	0.32
France	13	0.051	0.762	0.503	0.051	0.762	0.503
Germany	16	0.236	0.605	0.38	0.236	0.605	0.38
Greece	13	0	.	.	0	.	.
Hungary	7	0	.	.	0	.	.
Iceland	2	0.55	0.078	0.085	0.55	0.078	0.085
Ireland	2	0.384	0.357	0.318	0.384	0.357	0.318
Israel	6	0.006	0.687	0.688	0.006	0.687	0.688

Between and Within Country Inequality in Regional Well Being

Italy	21	0.003	0.885	0.476	0.003	0.885	0.476
Japan	10	0.001	0.639	0.322	0.001	0.639	0.322
Korea	7	0	0.825	0.483	0	0.825	0.483
Latvia	6	0	.	.	0	.	.
Lithuania	10	0	.	.	0	.	.
Luxembourg	1	0.709	0	.	0.709	0	.
Mexico	32	0	.	.	0	.	.
Netherlands	12	0.728	0.157	0.24	0.728	0.157	0.24
New Zealand	14	0.338	0.384	0.27	0.338	0.384	0.27
Norway	7	0.998	0.001	0.105	0.998	0.001	0.105
Poland	16	0	.	.	0	.	.
Portugal	7	0	0.857	0.465	0	0.857	0.465
Slovak R.	4	0	.	.	0	.	.
Slovenia	2	0	.	.	0	.	.
Spain	19	0.032	0.855	0.485	0.032	0.855	0.485
Sweden	8	0.925	0.054	0.212	0.925	0.054	0.212
Switzerland	7	0.352	0.363	0.264	0.352	0.363	0.264
Turkey	26	0	.	.	0	.	.
UK	12	0.297	0.574	0.418	0.297	0.574	0.418
US	51	0.585	0.358	0.454	0.585	0.358	0.454

Authors' elaboration on OECD (2016) data

Note: N=Number of observations; p = share of population; mean = average G = Gini coefficients; O = average overlapping.

4. References

1. Alesina, A., Rodrik, D. Distributive politics and economic growth. Quarterly Journal of Economics **104**, 465–490 (1994).
2. Bourguignon, F., & Morrisson, C. Inequality and development: the role of dualism. Journal of development economics **57(2)**, 233-257 (1998).
3. Coco, G., Lagravinese, R., & Resce, G. Beyond the weights: a multicriteria approach to evaluate inequality in education. The Journal of Economic Inequality **18(4)**, 469-489 (2020).
4. Costanza, R., Daly, L., Fioramonti, L., Giovannini, E., Kubiszewski, I., Mortensen, L. F., ... & Wilkinson, R. Modelling and measuring sustainable wellbeing in connection with the UN Sustainable Development Goals. Ecological Economics **130**, 350-355 (2016).
5. Costanza, R., Kubiszewski, I., Giovannini, E., Lovins, H., McGlade, J., Pickett, K. E., ... & Wilkinson, R. Development. Nature **505(7483)**, 283-285 (2014).
6. Decancq, K., & Schokkaert, E. Beyond GDP: Measuring social progress in Europe. Social Indicators Research, **126**, 21-55 (2016).
7. Döpke, J., Knabe, A., Lang, C., & Maschke, P. Multidimensional well-being and regional disparities in Europe. Journal of Common Market Studies **55**, 1026-1044 (2017).
8. Durand, M. The OECD Better Life Initiative: How's Life? and the Measurement of Well-Being. Review of Income and Wealth **61(1)**, 4-17 (2015).
9. Greco, S., Ishizaka, A., Matarazzo, B., & Torrisi, G. Stochastic multi-attribute acceptability analysis (SMAA): an application to the ranking of Italian regions. Regional Studies **52(4)**, 585-600 (2018).
10. Iammarino, S., Rodríguez-Pose, A., and Storper, M. Regional inequality in europe: evidence,theory and policy implications.Journal of economic geography **19(2),** 273–298 (2019).

11. Lagravinese, R., Liberati, P., & Resce, G. Exploring Health Outcomes by Stochastic Multicriteria Acceptability Analysis: An Application to Italian Regions. European Journal of Operational Research **274(3)**, 1168-1179 (2019).
12. Lahdelma, R., Hokkanen, J., Salminen, P. SMAA-stochastic multiobjective acceptability analysis. European Journal of Operational Research **106(1)**, 137-143 (1998).
13. Lahdelma, R., Salminen, P. SMAA-2: Stochastic multicriteria acceptability analysis for group decision making. Operations Research **49(3)**, 444-454 (2001).
14. Krugman, P. Increasing returns and economic geography. Journal of political economy **99(3)**, 483–499 (1991).
15. Kuznets, S. Economic growth and income inequality. American Economic Review **45(1)**, 1–28 (1955).
16. OECD. How's Life in Your Region? Measuring Regional and Local Well-being for Policy Making, OECD Publishing, Paris (2014).
17. OECD. OECD Regions at a Glance 2016, OECD Publishing, Paris (2016).
18. Peiró-Palomino, J. Regional well-being in the OECD. The Journal of Economic Inequality **17(2)**, 195-218 (2019).
19. Peiró-Palomino, J., & Picazo-Tadeo, A. J. OECD: One or many? Ranking countries with a composite well-being indicator. Social Indicators Research **139(3)**, 847-869 (2018).
20. Stiglitz, J. E., Sen, A., & Fitoussi, J. P. Report by the commission on the measurement of economic performance and social progress. Paris: Commission on the Measurement of Economic Performance and Social Progress (2010).

Equitable and sustainable well-being over time: a functional approach

Il benessere equo-sostenibile nel tempo: un approccio funzionale

Tonio Di Battista, Eugenia Nissi and Annalina Sarra

Abstract In recent years, studies on well-being have risen to prominence and it has been widely accepted that well-being should be considered a multidimensional phenomenon, beyond its economic feature. In vein of multidimensionality, Italy launched the Equitable and Sustainable well-being (BES) as the official framework for measuring well-being. Exploiting the availability of BES indicators over different sequential years, our research is aimed at computing well-being efficiencies of the Italian province capital cities. The procedure employed in this paper integrates the Malmquist DEA scores with the diversity profile to rank the cities. This joint approach has the benefit to be recast into a functional framework.

Abstract *Negli ultimi anni sono saliti alla ribalta gli studi sul benessere ed è stato ampiamente accettato che il benessere dovrebbe essere considerato un fenomeno multidimensionale, al di là della sua caratteristica economica. All'insegna della multidimensionalità, l'Italia ha lanciato il benessere equo e sostenibile (BES) come quadro ufficiale per misurare il benessere. Sfruttando la disponibilità di indicatori BES su diversi anni consecutivi, la nostra ricerca è finalizzata al calcolo delle efficienze dei capoluoghi di provincia italiani nel promuovere il benessere. La procedura utilizzata in questo lavoro integra i punteggi degli indici di Malmquist con il profilo di diversità per classificare le città. Questo approccio congiunto ha il vantaggio di poter essere ricollocato all'interno dell'analisi funzionale.*

Tonio Di Battista
University G. d' Annunzio of Chieti-Pescara, Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, Viale Pindaro, Pescara, e-mail: tonio.dibattista@unich.it

Eugenia Nissi
University G. d' Annunzio of Chieti-Pescara, Department of Economics e-mail: eugenia.nissi@unich.it

Annalina Sarra
University G. d' Annunzio of Chieti-Pescara, Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, Viale Pindaro, Pescara e-mail: annalina.sarra@unich.it

Key words: well-being, composite indicators, Malmquist index, Shannon-entropy, functional diversity

1 Introduction

Over the recent decades, the interest in the well-being measurement is constantly increased worldwide. In reviewing the recent trends on this topic,we acknowledge a more elevated scientific standard and a more rigor of approaches proposed for national and international comparisons of well-being. One major theme in this, has been the recognition of the limitations of macroeconomic indicators, Gross Domestic Product (GDP) in particular, as proxies for describing and measuring the factors affecting people's lives [8]. Accordingly, there has been a shift toward a conceptualization of well-being as a multidimensional phenomenon, as emerged in various arenas including the United Nations, the OECD, and numerous national, regional, and local governments [2, 12, 15]. Focusing on the Italian scenario, we consider, in this paper, the Equitable and Sustainable Well-Being Index, whose Italian acronym, used hereafter, is BES, elaborated by a joint initiative of the National Committee for Economy and Labour (CNEL) and the ISTAT, with the final aim of developing a collective definition of progress in the Italian society and producing a shared set of indicators of the most relevant economic, social and environmental domains. The theoretical framework adopted by Istat within the BES project can be deemed as the adjusted version, for the Italian context, of the conceptual model published by [9]. Alongside with the national experience, Istat has implemented the BES framework at local level and launched UrBES project to measure well-being at an urban level. Exploiting the availability of UrBES indicators over different sequential years, the core idea of this research is to compare over time the efficiencies of Italian provinces capital cities in producing equitable and sustainable well-being, by employing entropy based Malmquist indices [11] recast into a Functional Data Analysis (FDA) approach [14]. The remainder of this paper is structured as follows. In Section 2, we provide details of the theoretical background of the Malmquist based entropy procedure and its reformulation into a functional framework while in Section 3 we describe the urban well-being data.

2 Methodological framework

The analysis considers different stages. Firstly, we rely on Mazziotta-Pareto's method of penalties [16] to summarize the statistical indicators into the domains individuated by the UrBes project. Subsequently, within the Data Envelopment Analysis framework, we coupled Malmquist index with Shannon's entropy to represent the overall change in the efficiency of Italian provinces capital provinces in promoting well-being. Finally, the ranking of Italian Provinces capital cities is facilitated

through a reformulation of the entropy based Malmquist procedure where a family of diversity indices is considered. In what follows, we provide the basics of both procedures.

2.1 Entropy based Malmquist Productive Index (MPI)

Data Envelopment Analysis (DEA), firstly introduced by Charnes et al. [1], is a non parametric approach originally developed to measure the relative efficiency of Decision Making Units (DMUs) within production context characterized by multiple inputs and outputs. The literature review reveals that during the last decades the scope of DEA has broadened considerably, with successful applications to social indicators (see, among others, [3]). A proper extension of DEA to the field of composite indicators implies to consider a DEA model with only outputs. To avoid the inconsistencies that arise in a model without inputs, a single unitary input has to be included as in [10]. A very useful tool, in DEA, to calculate the relative performance of a DMU at different periods of time is the Malmquist Productivity Index (MPI). Suppose that there are n DMUs to be evaluated in lights of m input and s output. We denote with X_j and Y_j the input and output vector at DMU_j and θ it is a measure of technical efficiency of DMU_0. The MPI allows to calculate the relative performance at different periods of time and is a cross measure, which considers the production of a DMU as the efficient frontier built in the next or previous instant:

$$MPI_0 = \left[\frac{\theta^t x_0^t y_0^t}{\theta^t x_0^{t+1} y_0^{t+1}} \frac{\theta^{t+1} x_0^t y_0^t}{\theta^{t+1} x_0^{t+1} y_0^{t+1}} \right]^{\frac{1}{2}} \tag{1}$$

where $\theta^t x_0^t y_0^t$ and $\theta^t x_0^{t+1} y_0^{t+1}$, $\theta^{t+1} x_0^t y_0^t$ and $\theta^{t+1} x_0^{t+1} y_0^{t+1}$ are respectively the input oriented efficiency measures of DMU_0 at period t and t+1. MPI_0 measures the productivity change between periods t and t+1 [6]. The overall tendency in productivity changes of DMUs over time periods is traditionally obtained through the average of productivity indices of sequential times, implicitly assuming that all sectional indices equally affect the level of productivity. In this work, to eliminate the equal-weight effect, we take into account Fallahnejad's algorithm [7] who applies the Shannon's entropy to obtain more objective weights in aggregating the Malmquist productivity indices.

For the N DMU_s and the relative inputs and outputs for k+1 times $(t_0, t_1, ..., t_k)$, the Malmquist indices at two sequential times t and t+1 can be computed. They are denoted by $MPI_{jt}, \quad j = 1 \dots n, \quad t = 1, \dots k$ and summarised as in Table 1.

The weighted Malmquist productivity index for each DMUj ($WMPI_j$) is obtained via some established steps, as detailed below.

In **Step 1**, the matrix in Table 1 is normalized dividing each element of the column by the sum of column:

$$p_{jt} = \frac{MPI_{jt}}{\sum_{j=1}^{n} MPI_{jt}} \tag{2}$$

Table 1 MPI matrix

	MPI1	MPI2	....	MPIk
DMU_1	MPI_{11}	MPI_{12}	....	MPI_{1k}
DMU_2	MPI_{21}	MPI_{22}	....	MPI_{2k}
DMU_3	MPI_{31}	MPI_{32}	....	MPI_{3k}
....	....	....	....	....
....	....	....	....	....
DMU_n	MPI_{n1}	MPI_{n2}	....	MPI_{nk}

The above normalization allows to eliminate anomalies due to different measurement units and scales.

In **Step 2**, the entropy h_t for all normalized MPI is calculated as:

$$h_t = -h_0 \sum_{j=1}^{n} p_{jt} ln p_{ij} \tag{3}$$

Step 3 involves the computation of the degree of diversification, defined as:

$$d_t = 1 - h_t, \quad t = 1, \dots k \tag{4}$$

If the values of the productivity of the DMUs are close, the weight of a given year can be considered weak in the aggregating process.

In **Step 4** the degree of importance of MPI at time t is obtained by setting

$$w_t = \frac{d_t}{\sum_{s=1}^{k} d_s}, \quad t = 1, \dots k \tag{5}$$

where $\sum_{t=1}^{k} w_t = 1$.

In **step 5**, the weighted MPI is calculated as:

$$WMPI_j = \sum_{t=1}^{k} w_t MPI_{jt}, \quad j = 1, \dots n. \tag{6}$$

2.2 Functional Malmquist Productive Index (FMPI)

As shown in Section 2, the weighted productivity index for each DMUs has been obtained by exploiting the Shannon's entropy method. In our context, the Shannon's entropy can be regarded as a diversity measure of DMUs. Actually, different indices could have been taken into account to delineate such diversity which could have lead to different rankings. To overcome this limitation, a possible solution is to a consider a family of diversity indices, dependent upon a single continuous variable, that depict graphically a diversity profile. Formally, to evaluate the DMUs diversity, we consider the β diversity profile proposed by Patil and Taille [13] in the environ-

mental setting:

$$\Delta_\beta = \sum_{i=1}^{n} \frac{(1-p_i)^\beta}{\beta} p_i \quad \beta \geq -1 \tag{7}$$

In Eq.7, $\sum_{i=1}^{n} \frac{(1-p_i)^\beta}{\beta}$ can be interpreted as a measure of relevance of each DMUs for the change in productivity over time expressed by each MPI and p_i is defined as above. Note that Shannon's entropy is a particular case of β diversity profile, resulting when $\beta \to 0$. The benefit of using Δ_β instead of the single Shannon's entropy index, as suggested in the procedure reviewed in the previous section, is to rely on a larger spectrum of diversity measures, obtained varying β from -1 to 1. It follows that Δ_β is a convex curve that can be studied in a functional framework. In this work, our proposal is to rewrite the Steps 3-5 of the original Fallahnejad's procedure by replacing h_t with Δ_β. Specifically, in the Step 4, w_t is reformulated as $w_t(\beta) = \frac{d_t(\beta)}{\sum_{s=1}^{k} d_s(\beta)} \quad \forall \beta$, whereas in Step 5, $WMPI_j$ is consequently replaced by $WMPI_j(\beta) = \sum_{t=1}^{k} w_t(\beta) MPI_{jt} \quad \forall \beta$. Thus, for each DMUs we are able to obtain a functional weighted MPI. The comparative analysis of the weighted MPI curves and the consequently ranking of DMUs are achieved by means of functional tools (analysis of derivatives, radius of curvature and length of a curve), as suggested in [4].

3 Application to UrBes data

To meet the statistical information needs of local communities, Istat designed Bes at local level in cooperation with local authorities, investigating the specific information needs of Italian Municipalities, Provinces and Metropolitan Cities and tuning a shared theoretical framework [5]. Bes measures at local level maintain a high level of quality and consistency with the Bes indicators system and constantly follow the evolution of the Bes framework. The set of indicators, illustrating the 12 domains relevant for the measurement of well-being, is updated and illustrated annually in the Bes report. For the application of procedure illustrated above, we refer to the Ur-Bes framework which appraises well-being by a great deal of variables. In 2020, the set of indicators has been expanded to 152 (it was 130 in previous editions), with a deep revision that takes into account the transformations that have characterised Italian society in the last decade, including those linked to the spread of the COVID-19 pandemic. Owing to the data unavailability and missing values, our analysis is restricted to 103 Province capital cities and takes into account eight out of twelve domains of the original Ur-Bes dataset. In particular, we focus on the following domains: "Health", "Education and Training", "Work and Life Balance", "Economic well-being", "Social Relationships", "Security", "Landscape and Cultural Heritage", "Environment".

References

1. Charnes, A., Cooper, W., Rhodes, E.: Measuring the efficiency of decision making units. Eur. J. Oper. Res. **2**, 429-444 (1978)
2. Commission, E.: Gdp and beyond. measuring progress in a changing world. Communication, European Commission (2009)
3. Despotis, D. K.: A reassessment of the human development index via data envelopment analysis. J. Oper. Res. Soc. **56**(8), 969–980 (2005)
4. Di Battista, T., Fortuna, F., Maturo, F.: Environmental monitoring through functional biodiversity tools. Ecol. Indic. **60**, 237-247 (2016)
5. Istat: Le differenze territoriali di benessere:una lettura a livello provinciale. Istituto Nazionale di Statistica (2019)
6. Fare R., Grosskopf S., Norris M., Zhang Z.: Productivity growth, technical progress, and efficiency change in industrialized countries. Am. Econ. Rev. **84**(1), 66-83 (1994)
7. Fallahnejad, R.: Entropy based Malmquist Productivity Index in Data Envelopment Analysis. Int. J. Data Envel. Anal. **5**(4), 1425–1434 (2017)
8. Fleurbaey, M.: Beyond gdp: The quest for a measure of social welfare. J. Econ. Lit. **47**, 1029–1075 (2009)
9. Hall, J., Giovannini, E., Morrone, A., Ranuzzi, G.: A framework to measure the progress of societies. Oecd statistics working papers, OECD Publishing (2010)
10. Koopmans, T.: Analysis of production as an efficient combination of activities. In Koopmans, T. C., editor, Activity analysis of production and allocation, pp 33-97. New York: Wiley (1951)
11. Malmquist, S.: Index Numbers and Indifference Surfaces. Trab. Estad. **4**, 209-242 (1953)
12. OECD: Compendium of OECD well-being indicators. Paris: Oecd, OECD (2011)
13. Patil, G., Taillie, C.: Diversity as a concept and its measurement. J. Am. Stat. Assoc. **77**, 548–567 (1982)
14. Ramsay, J., Silverman, B.: Functional Data Analysis, second ed. Springer-Verlag, New York (2005)
15. Stiglitz, J., Sen, A., Fitoussi, J. P.: Report by the commission on the measurement of economic performance and social progress (2009) Paris.
16. Mazziotta, M., Pareto, A.: On a generalized non-compensatory composite index for measuring socio-economic phenomena. Soc. Indic. Res. **127**(3), 983–1003 (2016)

Session of free contributes SCL9 – *Health and Covid-19*
Chair: Maria Sole Pellegrino

Twitting about COVID-19: An application of Structural Topic Models to a sample of Italian tweets

Una Applicazione degli Structural Topic Models su un campione di tweets in lingua italiana sulla vaccinazione anti COVID-19

Niccolò Cao, Antonio Calcagnì and Livio Finos

Abstract During the COVID-19 pandemic, vaccination emerged as a burning issue in the Italian public discussion. In particular, social media were an important vehicle for spreading news, information, and opinions, both true and false, regarding health. In this contribution, we present an application of Structural Topic Model (STM) to a tweets-based corpus concerning the Italian public debate about COVID-19 vaccines. The aim is to detect the evolution of tweets-related topics characterizing the Italian public opinion about COVID-19 vaccination.

Abstract *Durante la pandemia di COVID-19, la vaccinazione è emersa come tema scottante nella discussione pubblica italiana. In particolare, i social media hanno rappresentato un importante mezzo di diffusione di notizie, informazioni e opinioni riguardo la vaccinazione, sia vere che false. In questo contributo, presentiamo un'applicazione dello Structural Topic Model (STM) a un corpus di tweets riguardanti il dibattito pubblico italiano in relazione ai vaccini contro il COVID-19. L'obiettivo è identificare l'evoluzione temporale dei topics associati ai tweets, che caratterizzano l'opinione pubblica italiana riguardo alla vaccinazione contro il COVID-19.*

Key words: Topic model, Structural Topic Model, tweets, vaccination, COVID-19.

Niccolò Cao
DPSS, University of Padua.
e-mail: niccolo.cao@studenti.unipd.it

Antonio Calcagnì
DPSS, University of Pauda.
e-mail: antonio.calcagni@unipd.it

Livio Finos
DPSS, University of Pauda.
e-mail: livio.finos@unipd.it

1 Introduction

Twitter is a social medium platform that is used by individuals, institutions, and public figures to communicate information and opinions. In the recent literature, the health topic has shown to have a growing trend in Twitter data-based researches (e.g. Public health, Infectious disease, Behavioral medicine, and Psychiatry) [4, 10]. In particular, the scientific literature about vaccination highlights the possibility that social media contents may influence significantly the users' opinions and behaviors [6]. Unfortunately, during the first stages of the COVID-19 vaccination campaign in Italy Twitter has been a receptacle of noteworthy amount of disinformation [7]. Moreover, Twitter platform tends to amplify reliable posts as well as unreliable posts with respect to COVID-19 news [3]. Within the Italian context, Twitter has demonstrated to be an useful gauge of public opinion regarding vaccination [13]. In this contribution, we present an exploration of the topics from Italian tweets concerning COVID-19 vaccination. The aim is to detect a kind of opinion trends about vaccination in the public discussions. To this end, we applied the Structural Topic Model (STM), which is specifically designed to analyze textual data with covariates [8]. In general, Topic models constitute a class of data mining techniques for categorical data, which have mainly been applied to textual data. The aim of these models is the extraction of semantic structures composed by words from a collection of documents [2]. The (hidden) topics can help researchers to examine in depth the idiosyncratic facets of a corpus.

2 Data and Methods

2.1 Data

The Twitter data used in the analyses were collected by [7], from December 20th, 2020 up to March 13th, 2021, with approximately 3 M tweets[1]. The keywords selected to download the tweets were related to the vaccination campaign (i.e. "vaccino", "novaccinoainovax", "iononsonounacavia", etc.). To this purpose, the `rtweet` package in R has been used [5].

2.2 Methods

Tweets data were analyzed by means of Structural Topic Model [8]. STM allows for the inclusion of document-specific covariates. The covariates can be applied to model topic proportions, or topic prevalence, ($\boldsymbol{\theta}_d$) over each document (d) or to influence the distribution of words within each k-th topic, or topic content, ($\boldsymbol{\beta}_k$).

[1] The original data are available to download at https://github.com/frapierri/VaccinItaly

The topic prevalence is assumed to be $\boldsymbol{\theta}_d \sim LogisticNormal(\boldsymbol{x}_d\gamma_k, \boldsymbol{\Sigma})$, where γ_k is the regression coefficient for the k-th topic and $\boldsymbol{\Sigma}$ is the covariance matrix. The topic content is the result of a multinomial probit regression in the form of $\boldsymbol{\beta}_k \propto \exp(\boldsymbol{m} + \boldsymbol{\kappa}_k^{(t)} + \boldsymbol{\kappa}_{\boldsymbol{y}_d}^{c} + \boldsymbol{\kappa}_{\boldsymbol{y}_d,k}^{(i)})$, where $\boldsymbol{\beta}_k$ is composed by the deviations of the k-th topic ($\boldsymbol{\kappa}_k^{(t)}$), selected covariate ($\boldsymbol{\kappa}_{\boldsymbol{y}_d}^{(c)}$), and topic-covariate interaction ($\boldsymbol{\kappa}_{\boldsymbol{y}_d,k}^{(i)}$) from the log-transformed overall words frequencies $\boldsymbol{m}$ [8].

With regards to the data analysis, the pre-processing procedure (i.e., text normalization, stop-word removal, tokenization, Document-Feature Matrix creation and sparsity reduction) was executed through the R libraries `TextWiller` [11] and `quanteda` [1]. The final number of tweets analyzed was about 2.874×10^6, with 3.127×10^7 word-tokens and 5558 word-types. We estimated the STM model using `stm` R library [9]. In order to identify non-linear changes of topic proportions over time, the topic prevalence was conditioned on the twitting-day using a B-spline with 10 degrees of freedom. The number of topics K was selected by running several STMs on the collected data. Finally, the model with $K = 25$ topics was chosen. Fig. 1 shows the results of the measures used to select the best models from a broad range of K. By contrast, Fig. 2 represents the comparison between the topics of the best fitted models in terms of semantic coherence and exclusivity.

3 Results and Discussion

Overall, the STM with $K = 25$ showed interpretable topics (there were few exceptions, for instance Topic 10, 12, and 19). Fig 3 shows a selection of meaningful

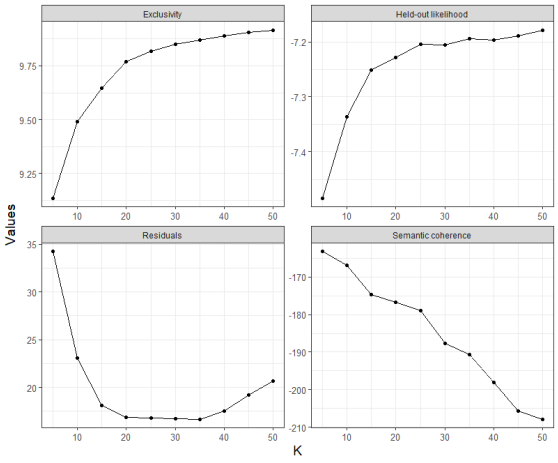


Fig. 1 Diagnostic measures used to select the number of topics in the data. Exclusivity is maximized when the most probable words occur within topic k more than within other topics; Held-out Likelihood refers to the probability of an held out set of data estimated by a previously trained model [14]; Residuals refers to the estimated residuals dispersion: more than 1 indicates that true K value is larger than the current [12]; Semantic coherence is maximized when the most probable words within one topic frequently co-occur together in documents [9].

topics according to the highest probability and FREX weightings[2]. In particular, Topic 23 reports news about the worst side effect of the first dose of vaccine; Topic 7 reports the updates about the Italian vaccination progress; Topic 11 concerns the general management of the pandemic situation; Topic 8 is associated with the foreign states news; Topic 3 regards the COVID-19 mutation and immunity; Topic 6 is focused on priority categories; Topic 14 is associated to fake news about vaccines; Topic 24 reports the issues in the vaccination plan in Lombardy. Fig. 4 reports the temporal trends of the selection of topics reported in Fig.3. The proportion of topics was in line with the main events behind the public discussion. For instance, Topic 24 shows the public attention on vaccination in Lombardy (we can notice a peak within the first month).

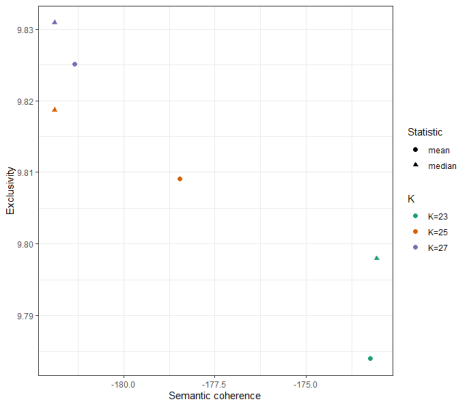


Fig. 2 Diagnostic measures for STM. Median and Mean are computed over all topics of each model. A model with higher semantic coherence and higher exclusivity has a better fit.

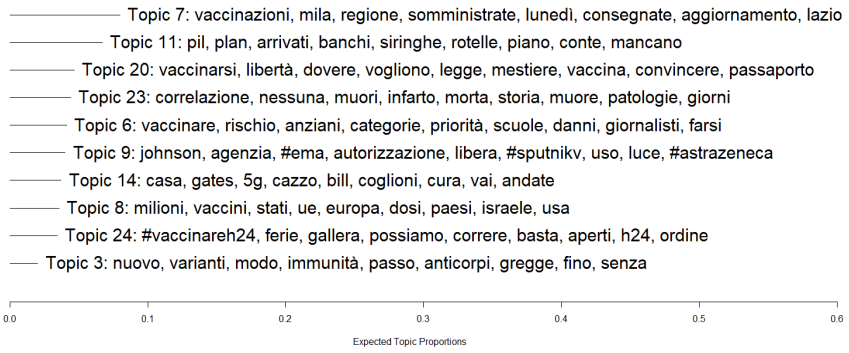


Fig. 3 Estimated topic proportions of STM ($K = 25$) with the highest probability words (Topics: 3, 6, 8) and FREX words (Topics: 7, 11, 14, 23, 24).

[2] FREX weighting refers to the weighted harmonic mean of the word rank in terms of exclusivity and frequency [9].

4 Conclusion

In the present short-paper, we explored the topics associated to the Italian public debate about the COVID-19 vaccination by means of topic modeling. All in all, the STM-based topics are able to reflect accurately the time trends of the daily news about Italian vaccination campaign and related events. Conversely, the topics can not be considered an indicator of noteworthy opinions or beliefs about vaccination, except for Topic 14 that summarize the main vaccination-associated fake news.

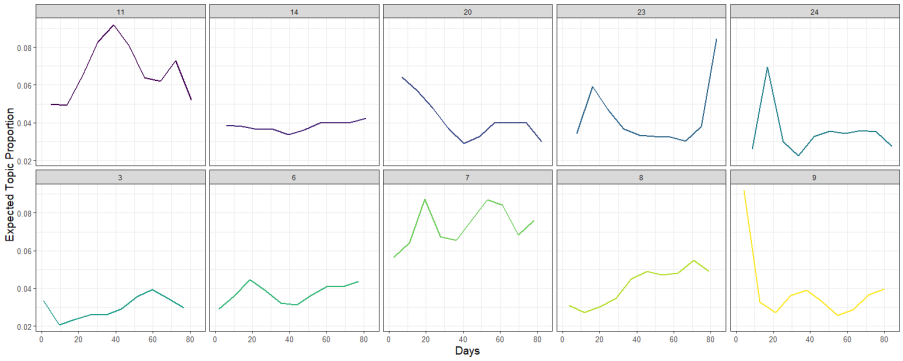


Fig. 4 Estimated proportion of topics over time. In the x-axis: 0 corresponds to December 20th, 2020, 40 to January 28th, 2021, and 80 to March 9th, 2021.

References

1. Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A.,: quanteda: An R package for the quantitative analysis of textual data. J. Open Source Softw. **3**, 774 (2018)
2. Blei, D., Carin, L., Dunson, D.:. IEEE Signal Proc. Meg. **27**, 55–65 (2010)
3. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 social media infodemic. Sci. Rep. **10**, 1–10 (2020)
4. Karami, A., Lundy, M., Webb, F., Dwivedi, Y.K.: Twitter and Research: A Systematic Literature Review Through Text Mining. IEEE Access **8**, 67698–67717 (2020)
5. Kearney, M.W.: rtweet: Collecting and analyzing Twitter data. J. Open Source Softw. **4**, 1829 (2019)
6. Ortiz, R.R., Smith, A., Coyne-Beasley, T.: A systematic literature review to examine the potential for social media to impact HPV vaccine uptake and awareness, knowledge, and attitudes about HPV and HPV vaccination. Hum. Vacc. Immunother. (2019) doi: 10.1080/21645515.2019.1581543
7. Pierri, F., Tocchetti, A., Corti, L., Di Giovanni, M., Pavanetto, S., Brambilla, M., Ceri, S.: VaccinItaly: monitoring Italian conversations around vaccines on Twitter and Facebook. Proc. Int. AAAI Conf. Web Soc. Media (2021)
8. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. **111**, 988–1003 (2016)
9. Roberts, M.E., Stewart, B.M., Tingley, D.,: Stm: An R package for structural topic models. J. of Stat. Softw. **91**, 1–40 (2019)
10. Sinnenberg, L., Buttenheim, A.M., Padrez, K., Mancheno, C., Ungar, L., Merchant, R.M.: Twitter as a tool for health research: a systematic review. Am. J. Public Health **107**, e1–e8 (2017)
11. Solari, D., Sciandra, A., Finos, L.: TextWiller: Collection of functions for text mining, specially devoted to the Italian language. Journal of Open Source Soft. (2019) doi: 10.21105/joss.01256
12. Taddy, M.: On estimation and selection for topic models. Proc. Artif. Int. Stat. **22**, 1184–1193 (2012)
13. Tavoschi, L., Quattrone, F., D'Andrea, E., Ducange, P., Vabanesi, M., Marcelloni, F., Lopalco, P.L.: Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. Hum. Vacc. Immunother. **16**, 1062–1069 (2020)
14. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. Proc. Mach. Learn. 1105–1112 (2009)

Assessing the quality of a health service through the risk profile number (RPN)

Valutare la qualità di un servizio sanitario attraverso l'indice del profilo di rischio.

Chiara Parretti, Riccardo Tartaglia, Giovanni Sbrana, Massimo Mandò, Samuele Pacchi

Abstract Healthcare systems are facing an important change that could design a new healthcare model through the adoption of digital technologies for the remote management of patients. There are still many aspects to be investigated, especially for the Italian Healthcare System. The adoption of tools based on (proactive) risk analysis, in the pre-implementation phases, can drive the development of technologies closer to the users, ensuring a fairer and safer healthcare.

Abstract *I sistemi sanitari stanno affrontando un cambiamento importante che potrebbe disegnare un nuovo modello di sanità attraverso l'adozione di tecnologie digitali per la gestione da remote dei pazienti. Gli aspetti da analizzare, soprattutto per il Sistema sanitario italiano, sono ancora molti. L'adozione di strumenti basati sull'analisi del rischio, nelle fasi pre implemntativa, possono guidare, lo sviluppo di tecnologie sempre più vicine agli utenti garantendo una sanità più equa e sicura*

Key words: Risk Assessment, FMEA, Health Systems, Telemedicine

[1] Chiara Parretti, Guglielmo Marconi University, Department of Engineering c.parretti@unimarconi.it

Riccardo Tartaglia, Guglielmo Marconi University, Department of Engineering, ri.tartaglia@unimarconi.it

Giovanni Sbrana, South East Tuscany Healtcare Agency, Department of Eemergency, giovanni.sbrana@uslsudest.toscana.it

Massimo Mandò, South East Tuscany Healtcare Agency, Department of Eemergency, massimo.mando@uslsudest.toscana.it

Samuele Pacchi, South East Tuscany Healtcare Agency, Department of Eemergency, samuele.pacchi@uslsudest.toscana.it

1 Introduction

Health care systems are currently facing a major challenge towards change and digitization in order to sustainably respond to rising costs due to multiple factors including the increase in chronic diseases and an aging population. [4]
The deployment of remote health monitoring systems, in an increasingly digitized environment, enables the management of various diseases and provides health care providers with reliable, real-time data to make clinical decisions faster, with a greater degree of reliability. During the Covid-19 crisis, the diffusion of digital health technologies increased exponentially in all countries as a response to social distancing measures, and the European Commission funded large sums to EU countries for the development of these tools. [3] .
Concerning the situation of the Italian Health System, there are still many aspects to be considered for an extensive implementation (accountability, human factors, usability, privacy and security, real effectiveness for improving health care, etc.)
Digital technologies for health offer the opportunity to develop a new model, where distance and the need for distancing are not a barrier, but an opportunity to use time more efficiently and reduce risks [8].

1.1 The Telemedicine

Telemedicine is one of the possible solutions to make healthcare services more efficient and optimize the time that medical personnel spend caring for patients. [13]. For the sake of clarity, it should be mentioned that the term telemedicine refers to the secure transmission of medical information and data in a variety of forms: text, sound, images, or other forms necessary for the prevention, diagnosis, treatment, and subsequent follow-up of patients. Under the term of telemedicine are grouped a series of services that do not only concern the direct relationship between doctor and patient, but embrace a much broader sphere of services ranging from the exchange of information between health care personnel, to the training of providers. [7] (see fig. 1) [6]

Typology		Scope	Users	
Specialist telemedicine	tele-visit	healthcare	can be used for acute, chronic, and post-acute disease management	active patient presence
	tele-consultation			patient absence
	telehealth coverage			real-time patient presence
Telemedicine		healthcare	mostly used for chronic diseases	active patient presence
Telecare		socio-health	aimed at frail elderly and disabled people	

Figure 1: telemedicine areas

The potential pool of users is extremely wide, in fact, there are more than 200 million people in Europe and the United States who suffer from one or more diseases in which telemedicine would be an excellent solution from a medical point of view, such as the management of chronic diseases. Currently in our country it is estimated that 40% of the population is affected by chronic disease, ie about 24 million Italians, a figure that is expected to grow, (see Table 1). Total spending in Italy has reached 66.7 billion. It is estimated that this figure will reach the threshold of 70.7 billion euros in 2028 [11]

Table 1: Projection of the number of people with chronic disease through 2038

Pathology type	***2017***	***2028***	***2038***
people with at least one chronic disease	21.040	25.233	25.289
people with at least two chronic diseases	12.578	13.907	14.673
diabetes	3.411	3.634	3.908
hypertension	10.702	11.846	12.523
bronchitis chronic	3.553	3.731	3.856
arthrosis /arthritis	9.723	10.803	11.506
osteoporosis	4.772	5.279	5.757
heart disease	2.499	2.689	2.926
allergies	6.428	6.313	5.940
nervous disturbances	2.732	2.925	2.978
ulcers of the gastric tract	1.435	1.586	1.611

2 Case Study

In this paper we present the analysis carried out in collaboration with the South East Tuscany Healtcare Agency on the service of remote monitoring of Covid 19 patients. The aim of the investigation was to identify, through tools such as FMEA and process analysis, critical issues and areas for improvement to be implemented in a service developed in an emergency. The analysis was conducted to optimize the monitoring process and verifying its reliability and efficiency in order to use it in other emergency situations or, in the medium to long term, to implement it for the management of chronic patients. In fact, the South East Tuscany Healtcare Agency, due to its characteristics, is an optimal candidate for the use of telemedicine for the ordinary management of chronic patients. It has an important territorial extension, compared to other health companies in Tuscany, with a low population density due to the territorial characteristics of the area and an important number of elderly population. The hospital resources are also concentrated in the 3 main cities: Siena, Arezzo and Grosseto which in turn have different information systems that cannot be interpolated between them.

2.1 Description of the monitoring process

The device used for monitoring is a small wearable composed of 3 main sensors that through a Bluetooth connection transfers data to an App installed on the patient's phone. The cell phone in turn sends the data to an operational monitoring center that can view the readings in real time. This allows physicians and operators to manage multiple patients at the same time, without the need to go on site if not for a real need for further investigation or hospitalization. Through remote monitoring, patients are managed in their own home environment, avoiding crowding emergency rooms, and greatly limiting the exposure of medical personnel to the risk of infection.

2.2 Fmea

The analysis carried out made particular use of FMEA being a widely used methodology for the analysis of systems and processes in healthcare. [12]. The FMEA is a methodology that supports the analysis of critical aspects of a process allowing to proactively identifying its vulnerability. It supports the identification of improvement actions based on risk assessment, through the definition of a Risk Priority Number (RPN). The RPN is defined through three parameters: Probability, Detectability and Severity [2] [10].
In our case, the survey was conducted to highlight the areas for improvement in the process under consideration and to develop solutions necessary to eliminate some of the problems that emerged in the emergency phase.
The analysis conducted saw the involvement of all the relevant figures within the patient monitoring process. In the project team, there were a doctor of emergency medicine, a nurse of the 112 operations center, a computer expert of the South East Tuscany Healtcare Agency, an engineer of the device manufacturing company, an expert of clinical risk and a management engineer.

3 Result

Through the analysis carried out, the necessary improvement actions have been identified and prioritized through the RPN. In particular, it emerged that due to a high rate of change of doctors employed in the USCA (Special Continuity of Care Units) it was necessary to develop checklists to be delivered to them that would guide them in the practical operations of installation and alignment of the device to the operations center.
The process of privacy management was particularly critical. In fact, the regulations on the use and management of health data are particularly stringent in our country

and in particular in Europe [1] [5]. For this reason, the management of privacy risk is an important constraint in the operation of tele monitoring and telemedicine in its entirety. Another particularly critical aspect is the impossibility for doctors to access the patient's Electronic Health File, if the patient has not previously activated or authorized it. This aspect exposes patients to a high risk of not being correctly framed from a clinical point of view, producing negative consequences in the therapeutic framework for the management of the disease (Table 2).

Table 2: summary of the aspects analyzed that were found to be particularly critical according to ipr

Process phases	***RPN***	***Main criticalities***
Instrumental evaluation	265	Inability to correctly assess the patient's clinical situation
Clinical evaluation of the patient	278	Incorrect diagnostic classification
Care plan evaluation	258	Incorrect therapeutic framework
Technology compatibility assessment	390	Inability to carry out monitoring, a problem present especially in elderly patients
Device registration	307	Blocked procedure
Device activation	453	Impossible to activate the device
Device/center alignment	237	Data don't reach the operating center
Data download from portal	251	The file is not transferable to the General Practitioner

From the analysis, however, interesting aspects also emerge to support the spread of tele monitoring, in particular:

- The data collected through the device are accurate and reliable and can be used to make clinical decisions in a short time because they are immediately available.
- The patient has the possibility to stay at home, reducing the risks of proximity and maintaining their habits, an extremely important aspect especially for frail or elderly patients.
- Data available in real time allow operators to identify false positives, avoiding unnecessary dispatch of personnel and limiting unnecessary access to hospital facilities.
- Operators have the ability to remotely test patients and verify results in real time
- Data flows continuously, there is no need to call patients or send a doctor to their homes to check their health status
- Using a Bluetooth connection the parameters are immediately available, no additional infrastructure is required.

4 Discussion and conclusion

The technique used has some limitations that must be considered during the analysis. In particular, operator experience plays an important role in the definition of scores. The RPN does not have statistically representative validity, but is defined through qualitative data. Therefore, the analysis must be supplemented with retrospective data to have predictive value. The results of the analysis remain valid as long as the process under consideration and the team participating in the analysis are not changed [12].

However, the adoption of assessment tools based on risk quantification, such as those used in our case study, allows the identification of corrective actions to be taken to ensure optimal quality of care, despite the limitations described above. In addition, it should be noted that if this type of investigation is carried out on processes that have not yet been developed, it is possible to identify in advance the most at-risk stages of the process. Such a mechanism has the great advantage of defining critical areas of the system in advance, making it possible to identify and apply the necessary adaptive interventions before problems arise [9].

The objective of our study is to improve the implementation of devices for remote monitoring of patients; these tools in fact if well used have proved very useful for home care

References

1. Enisa, (2017) Handbook on Security of Personal Data Processing
2. EN IEC 60812: 2018-10 Failure model and effects analysis
3. European Commission. Recovery and Resilience Facility https://ec.europa.eu/info/business-economy-euro/recovery-coronavirus/recovery-and-resilience-facility_en (2021)
4. Flott, K., Fontana, G., Dhingra-Kumar, N., Yu, A., Durkin, M., & Darzi, A.: Health care must mean safe care: enshrining patient safety in global health. The Lancet, 389(10076), 1279-1281(2017) doi: 10.1016/S0140-6736(17)30868-1
5. General Data Protection Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council, Official Journal of the European Union
6. Linee d'indirizzo nazionali sulla Telemedicina. Ministero della Salute (2019)
7. Klaassen, B., van Beijnum, B. J., & Hermens, H. J.: Usability in telemedicine systems - A literature survey. In International journal of medical informatics **93,** 57-69, (2016) doi: 10.1016/j.ijmedinf.2016.06.004
8. Nouri, S., Khoong, E. C., Lyles, C. R., & Karliner, L.: Addressing equity in telemedicine for chronic disease management during the Covid-19 pandemic. NEJM Catalyst Innovations in Care Delivery, 1(3) (2020) doi: 10.1056/CAT.20.0123
9. Parretti, C., Pourabbas, E., Rolli, F., Pecoraro, F., & Citti, P.: Robust Assessment in Transnational Healthcare Systems. In IOP Conference Series: Materials Science and Engineering, **1174**(1), 12-15. IOP Publishing (2021) doi:10.1088/1757-899X/1174/1/012015
10. QI Essentials Toolkit: Failure Modes and Effects Analysis (FMEA). Institute Healthcare Improvement

Assessing the quality of a health service through the risk index
11. Rapporto Osservasalute 2019 Stato di salute e qualità dell'assistenza nelle regioni italiane
12. Toccafondi G, Dagliana G, Fineschi V, Frati P, Tartaglia R.: Proactive Risk Assessment through FMEA of Home Parenteral Nutrition Care Processes: A Survey Analysis. Curr Pharm Biotechnol.; **22(3),** 433-441(2021) doi: 10.2174/1389201021666200612171943.
13. Tuckson, R. V., Edmunds, M., & Hodgkins, M. L.: Telehealth. New England Journal of Medicine, **377(16)**, 1585-1592, (2017) doi: 10.1056/NEJMsr1503323

THE INSURANCE PREMIUM STRUCTURE FOR A COVID-19 INSURANCE POLICY.

Struttura di un premio assicurativo per una polizza COVID-19.

Giovanna Di Lorenzo, Girolamo Franchetti and Massimiliano Politano

Abstract In the context of the Sars-CoV-2 virus pandemic, this paper deals with the analytical framework that involves a stochastic model to describe the probability of contagion and, therefore, of its outcome for a subject; subsequently, an actuarial model for an insurance policy against the risk of contracting the virus is proposed and the quantification of the related premium. It is assumed that the insurance coverage lasts for one year and that during the coverage it could happen the infection. The theoretical distribution of the contagion probability is of geometric type, in which every coverage day is a Bernoulli distribution of infection event. Four outcomes of the infection are considered below: hospitalization in home isolation, in hospital Medic Area, in intensive care and, finally, death. The Gamma distribution is taken into account as the theoretical distribution of number of days for each trajectory of recovery regarding the outcome of the infection, whereas for the outcome of death a lump-sum payment is defined to be paid as a single solution. A payment variable will be obtained whose mathematical expectation is the expected value of the expected benefits, assuming that in the event of death it remains the capital value. For each day of coverage, the expected payment is calculated and then weighted by the probability of infection on that given day; then, the expected payment is discounted to the effective date of coverage and, finally, it is calculated

[1]Giovanna Di Lorenzo, University of Naples "Federico II"; email:giodilor@unina.it

Girolamo Franchetti, University of Naples "Fe; email:gi.franchetti@studenti.unina.it

Politano Massimiliano*, University of Naples "Federico II"; email:politano@unina.it
*Corresponding Author

the fair premium of the policy. For this paper it is used the software R for statistical evaluation

Abstract *Nel contesto della pandemia Sars-CoV-2, questo lavoro affronta il quadro analitico che prevede un modello stocastico per descrivere la probabilità di infezione e, quindi, del suo esito per un soggetto; successivamente viene realizzato un modello attuariale per una polizza assicurativa contro il rischio di contrarre il virus e la quantificazione del relativo premio. Si ipotizza che la copertura assicurativa abbia una durata di un anno e che durante la copertura possa verificarsi il contagio. La distribuzione teorica della probabilità di contagio è di tipo geometrico, in cui ogni giorno di copertura è una distribuzione Bernoulliana dell'evento di infezione. Di seguito vengono considerati quattro esiti del contagio: il ricovero in isolamento domiciliare, in Area Medica ospedaliera, in terapia intensiva e, infine, il decesso. La distribuzione Gamma viene presa in considerazione come distribuzione teorica del numero di giorni per ogni traiettoria di ricovero in relazione all'esito della guarigione, mentre per l'esito del decesso si definisce un'indennità una tantum da erogare in un'unica soluzione. Si otterrà una variabile di pagamento la cui aspettativa matematica è il valore atteso dei benefici attesi, assumendo che in caso di morte rimanga il valore del capitale. Per ogni giorno di copertura viene calcolato il pagamento previsto e poi pesato per la probabilità di infezione in quel dato giorno; quindi, il pagamento previsto viene attualizzato alla data di decorrenza della copertura e, infine, viene calcolato il premio equo della polizza. Per questo lavoro viene utilizzato il software R per la valutazione statistica.*

Key words: Sars-Cov2, Covid-19 insurance, fair premium, vaccinated ad unvaccinated risk profile

1 The Model

The model is composed by two different sections strictly linked each other:
1 A stochastic model for the description of COVID-19 epidemic dynamics
2 An actuarial model to quantify a fair premium of a COVID-19 insurance policy.

1.1 Stochastic model

The stochastic model takes in account the data presented by the Instituto Superiore della Sanità (ISS) in his weekly report about the COVID-19 epidemics during the observation period from 14/07/2021 to 17/11/2021. It is calculated the probability of infection using the classic approach, taking in account the entire reference population, and the probability of a specific outcome of the infection itself taking in account infected people as the reference population:

Table 1: Infection and related outcomes probabilities

Non infection	*Infection*	*Home quarantine*	*Hospitalization*	*Intensive care*	*Death*

The Insurance premium structure for a Covid-19 insurance policy

$$p^S = \frac{S}{N} \quad p^I = \frac{I}{N} \quad p^C = \frac{C}{I} \quad p^O = \frac{O}{I} \quad p^{TI} = \frac{TI}{I} \quad p^D = \frac{D}{I}$$

Where

$p^S + p^I = 1$

$p^C + p^O + p^{TI} + p^D = 1$

Where *S* stands for people who are susceptible to the infection but not infected. Calculating these probabilities for every observation data it is possible to see the time series of this probabilities.

1.2 Actuarial model

Using the probabilities from the stochastic model seen before it is possible to quantify the fair premium of a COVID-19 insurance policy.

It is assumed that the policy coverage period is of 365 days and that the infection is represented by a bernoulli variable for every day of this period. Then, it is calculated the probability of being infected for every single day assuming that in the previous days the infection event has not happened.

$$\mathcal{B}(p^I) = \begin{cases} 0 & 1-p^I \\ 1 & p^I \end{cases}$$

$$p^I(t) = \mathcal{G}(p^I) = (1-p^I)^{t-1} * p^I$$

So, it is used a geometric distribution for representing this probability for every coverage day and, considering every single observation data, it is possible to obtain the dynamic of this probability during the observation period.

For every infection outcome (except death) it is calculated the expected benefit calculating the expected value of recovery days multiplied for the benefit paid every single recovery day. For death case it is paid a lump sum capital.

$$E[P] = E[rC] * p^C + E[rO] * p^O + E[rTI] * p^{TI} + 300 * p^D$$

Table 2: Benefit related to the infection outcomes (first three paid for every recovery day)

Home quarantine	*Hospitalization*	*Intensive care*	*Death*
1	1.5	2	300

It is used the probabilities related to the four possible outcomes to calculate the expected benefit of the policy by calculating the expected value of the expected benefits.

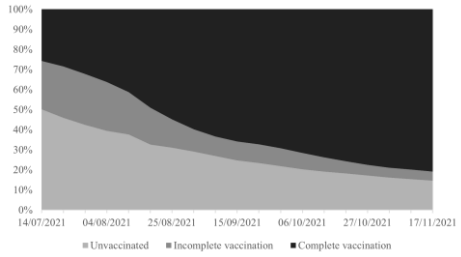
Finally, the expected payment is assumed the same for every day of the coverage period, weighted by the related infection probability and it is discounted at a technical interest rate of 1%.

$$Fair\ premium = \sum_{t=1}^{365} V_0(t) = \sum_{t=1}^{365} (1+i)^{-\frac{t}{365}} * p^I(t) * E[P]$$

3 Data

It is used the data present in the weekly report about COVID-19 epidemic published by the Istututo Superiore della Sanità (ISS) about the infection and the related outcomes. This data are shown for three vaccination status: Unvaccinated, incomplete vaccination, complete vaccination.

Graph 1: Population partition by vaccination status



Regarding the recovery days, related to non-death outcomes, it is used the mean of the hospitalization and intensive care days presented by Azienda Sanitaria Toscana (ARS Toscana) and the mean of recovery days related to the home quarantine presented by the Health Ministry. So, the gamma distribution is considered as the theoretical distribution of number of days for each trajectory of recovery regarding the outcome of the infection with shape parameter equal to 2.

Graph 2: Theoretical gamma distribution of recovery days

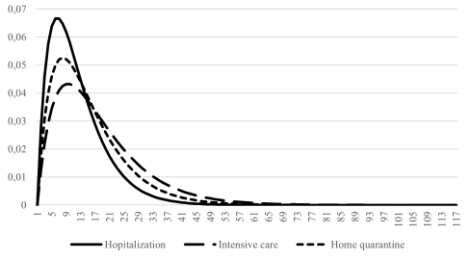


Table 3: Means of the recovery days

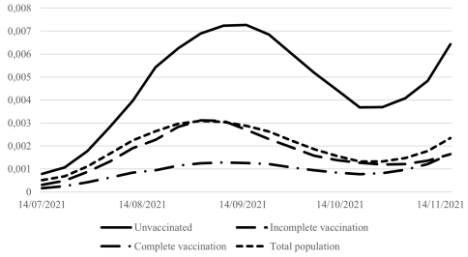
Home quarantine	***Hospitalization***	***Intensive care***
14	11	17.3

4 Analysis

In the analysis it is computed the time series of infection probability and the time series of the expected benefits by vaccination status. Then, it is calculated the fair premium for each class of people differentiated by vaccination status.

The Insurance premium structure for a Covid-19 insurance policy

Graph 3: Time series of infection probability by vaccination

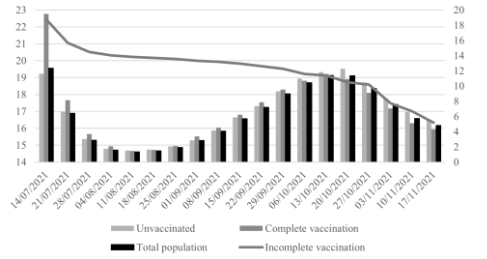


So, it is possible to see that during August and September the probability of infection increased due to the wave of infections and the same is happening during November.
Another statement is that the infection probability is higher than the other vaccination status and higher than the probability calculated on entire population. This data shows that the class of unvaccinated is the one with the highest risk about infection event.
Regarding the expected benefits related the infection outcome, they are shown in the graph except the lump sum capital in death case. The highest is the intensive care due to the mean of recovery days and the payment in each one that are the highest. The lowest is the home quarantine for the same reason presented before but the difference that in this case they are the lowest.

Table 3: Expected benefits for the infection outcome

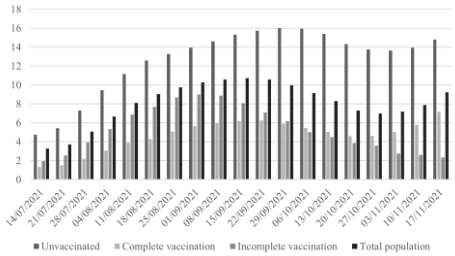
Home quarantine	***Hospitalization***	***Intensive care***
14,02368661	16,54542718	34,03564112

Graph 4: Expected payment of the policy by vaccination



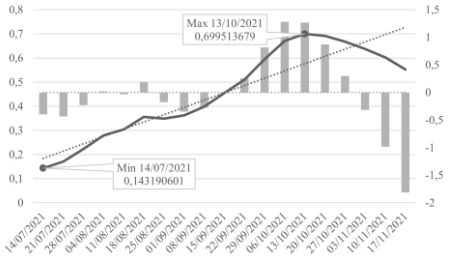
It is possible to state that during the wave of the infections, most of them are people that have the home quarantine as the infection outcome. So, this brings to see a reduction of expected payment during the wave, because it is more expected that it is paid the lowest benefit. Only by 20/10/2021 the expected benefit for the unvaccinated is higher than the expected for the entire population. Applying the actuarial model, it is possible to see that the class of unvaccinated is the one with the highest fair premium compared to the other vaccination status, confirming that this is the class of the highest risk related to the infection event and to the possible benefit. It is possible to see that the premium increases during the wave, as expected, and decreases outside it. Observing this result, it is possible to state that the vaccination status has a relevant impact on the quantification of fair premium.

Graph 5: Time series of the fair premium by vaccination



In fact, people who is unvaccinated has a higher risk to get infected and have a higher risk to be in more serious infection outcome.

Graph 6: Time series of the fair premium gap between unvaccinated people and total



In conclusion, if an insurance company sells a COVID-19 insurance policy to the entire population has a relevant risk to incur in losses due to the serious mismatch between the fair premium of the policy and the one that unvaccinated people should pay that is much higher that the one of the insurance policy offered to the entire population.
So, there is the necessity to apply a safety loading on the probability of infection and on the expected payment to cover the risk that the company could provide a payment higher that expected in the offered policy if the insured person who incurs in the infection has the unvaccinated status.

References

1. F. Gemmi, F., L. Bachini, L., S. Forni S.: I ricoveri per Covid-19 in Toscana. Aggiornamento "terza ondata" (2021). www.ars.toscana.it/2-articoli/4649-ricoveri-per-covid-19-in-toscana-aggiornamento-terza-ondata.html
2. Greenwood, P. E., Gordillo, L. F.: Stochastic Epidemic Modeling. In: Chowell, G., Hyman, J. M., Bettencourt, L. M. A., Castillo-Chavez, C. (eds) Mathematical and Statistical Estimation Approaches in Epidemiology, pp. 31-52. Springer (2009).
3. He, S., Hang, S., Rong, L.: A discrete stochastic model of the COVID-19 outbreak: Forecast and control. Mathematical Biosciences and Engineering **17**, 2792-2804 (2020)
4. Kermack, W.O., McKendrick, A. G.: A Contribution to the Mathematical Theory of Epidemics. Bulletin of Mathematical Biology **53**, 33-55 (1991).
5. Istituto Superiore di Sanità: Aggiornamento dei dati: i bollettini della sorveglianza integrata COVID-19 in Italia (2021). www.epicentro.iss.it/coronavirus/aggiornamenti
6. Pitacco, E.: Elementi di matematica delle assicurazioni. Lint Editoriale. Trieste (2016)
7. Zhou, Y., Ma, Z.: A Discrete Epidemic Model for SARS Transmission and Control in China. Mathematical and Computer Modelling **40** (13), 1491-1506 (2004).

Session of solicited contributes SS22 – *Statistics, culture and tourism*
Organizer and Chair: Marica Manisera

How data can influence the promotion and consumption of a cultural experience

Come i dati possono influenzare la promozione e la fruizione dell'esperienza culturale

Piera Cristiani

This focus on the improvement of cultural consumption has two points: the first looks at the communication job and the development of the relationship with the press; the second insists on the necessity for cultural institutions to get data used to improve the experience of people during the time of visit.

Questo testo vuole essere una riflessione sulle possibilità della fruizione culturale basato su due punti: il primo si snoda dal lavoro di comunicazione e lo sviluppo delle relazioni con il mondo della stampa; il secondo insiste sulla necessità delle istituzioni culturali di elaborare alcuni dati che possono essere utilizzati per incrementare l'esperienza dei visitatori durante il tempo di visita.

Key words: cultural experience, data, cultural product

[1] Piera Cristiani, Communication specialist and PR; email: info@pieracristiani.com

Find the difference: the cultural product

The purpose of art is washing the dust of daily life off our souls.
Pablo Picasso

Is cultural promotion really considered as marketing? The products created in the cultural area have different dynamics when it comes to involving people, starting from their main characteristic of being immaterial. First, culture looks beyond property and what is collected in the cultural field is related to the intellectual and spiritual heritage, apparently not in touch with the material life of objects. Secondly, there is no physical condition of need related to culture and the budget invested in culture has a very different weight.
Culture has been acknowledged as a driver of economic development and as a contribution to economic growth, general well-being, social cohesion and the whole sector is also an excellent way to support diversity. Statistics on culture help to answer questions on its impact on the whole economy as well as enable to draw the picture of societal aspects of culture, such as how many people participate in cultural events, how much households or governments spend on culture, and many more.
What is currently defined as cultural product?
Mostly, we use this definition for experiences that connect people to their intellect.
The cultural product could be declined in a lot of ways, and it could involve many disciplines, but what is important is to focus on the immaterial part of this offering: going to the theatre, visiting an art exhibition, experiencing a performance, going to a concert, attending a special lecture, watching a movie are not business-related activities, they are mostly driven by a peculiar mindset.
Every time people are involved in a cultural experience, the analytic part of the brain makes space for imagination and for any emotion that can be unleashed in contact with the experience itself.
At the same time, this kind of activity enhances the process of excitement before the show in a theatre or the curiosity of starting along a museum's path. These singular sensations become bricks of people's identity, taste and personality and the choice of being in touch with the emotional part of one's emotions is a guarantee of an open-minded attitude that will translate into other aspects of a person's life.

1.1 How to promote a cultural product

The current information revolution is a cultural revolution and a social revolution, a thoroughgoing technological revolution that involves not just

information, but labor, leisure, entertainment, communication, education, culture and thus is part of a major cultural and social shift
Bill Gates

The promotion of cultural products engages different targets on an immaterial experience, so it must follow its own rules.
After deciding which objective to reach, it is fundamental to create a solid and proper strategy to build the right audience and that's where the traditional rules of marketing become useful: find a big goal, submit smaller objectives, study the target, choose the way to achieve them.
When the goal is to build an audience, it's hard to focus on just one range of public because people connected to the cultural sector are vastly different, so the messages must be varied and divided into goal-driven activities.
The proper feature of being an untraditional product makes the cultural experience demand a deep focus on content: if the promotion is not aligned with the common understanding of each theme, or the standard of the language is not appropriate, some problems of acceptance by the audience might emerge due to the complexity of one of the most difficult sectors to work with.
These are just some key points that prove why audiences are so important.

1.2 Mailchimp

I think it's very important to have a feedback loop, where you're constantly thinking about what you've done and how you could be doing it better. I think that's the single best piece of advice: constantly think about how you could be doing things better and questioning yourself.
Elon Musk

The latest platforms for sending emails have definitely improved the email marketing. Tools like Mailchimp or MailUp have changed the mailing activities, not only for direct sales along with the growth of online shopping, but also for a press office or for a content newsletter service.
Obviously, the pillars of any PR activity are agenda, reputation, contacts and the quality standard of the proposal, but the more people you can reach with just one click, the better and more time saving it all gets.
The real advantage of these platforms is the insight: every campaign has its own analysis with the percentage of email opening, overall number of people along with names, how many times each email has been read and each link has been clicked on. Each unsubscription gets noticed and there is a weekly summary of activities, trends of results, locations in the world from where emails have been opened and, sometimes, the range of age.

What is really useful is the possibility to get close to the audiences and the more contacts are divided into segments, the more efficient the analysis phase gets. This way, it is possible to understand how many people are interested in the content, thanks to the rate of click and opening of each email received, and this is a key point for engaging into a deeper and more straightforward contact with the audience.

This powerful instrument is important to trace and build a proper recollection of journalists, who likes who, who's qualified for what, who never answers, who prefers photography, etc. This big baggage of small details is fundamental to consolidate relationships, and relationships are the foundation of this job.

1.3 Social Networks

Big Data is like teenage sex: everyone talks about it, nobody really knows how to do, everyone thinks everyone else is doing it, so everyone claims they are doing it.
Dan Ariely

Social Networks represent a forefront in the relationship with the audience, but they need to be managed as business tools. They must reflect the message as planned, but they must seem spontaneous and not too inward looking: it's hard work! After a study aimed to decide which media is best suited, it is fundamental to distinguish each and every content, draw an editorial calendar and start with the publication.

The most interesting part starts then, when it's time for an activity that must not be ignored: insights.

Having an online activity without studying its performance could be a waste of time, because much can be learned about the audience: age, geographical area, trends, kind of job, tastes.

Why do we need this data?

First of all, we could use the online and offline presence to study all of the numbers and put more effort on what is working and, on the other hand, understand what doesn't work and why. Social media is drawing public profiles closer and closer to their audience, and they can interact by making questions, asking preferences, picking between alternatives, react with different emoticons: every answer captured from the audience is a precious element to go deeper into connection and make actions more goal-oriented.

The most meaningful data goes beyond followers, visualizations or likes, but also beyond the distribution of likes, for example.

Some fundamentals key elements to analyze are the distributions of likes through geographical areas or the different ages of people reacting to different proposals:

what we must know is how to reach the people we are interested in, so every step that makes it closer is a small victory.

2 How data could improve the experience in cultural area

Without tradition art is a flock of sheep without a shepherd. Without innovation it is a corpse.
Winston Churchill [4]

People enter a cultural experience with an expectation, sometimes high and vibrant, and other times they might be skeptical or simply unaware. On the other side, at the end of the performance, something has changed.
The hard part for a cultural institution is to understand if its programs work, because this could influence its future in terms of credibility and viability of upcoming projects.

2.1 A dialogue with the visitors

Art is what you can get away with
Andy Warhol

The question is how to have a productive relationship with visitors.
The plan is to be effective, the same way you'd want a friendship to grow: you see each other, you talk, you share experiences and your confidence grows over time to time.
It's possible to retain an audience by taking care of each detail of the experience, but and in order to do that you need answers. One of the most common techniques it's the survey at the end of the experience: a simple form to fill in as a way to discover the level of satisfaction.
The questions should cover the whole experience: quality, content, experience, kindness of personnel, access to the area, etc.
In some cases, the survey can be submitted to a specific membership group, more attached to the institution and therefore easier to reach.
If the inquiry concerns an educational area, it is possible to involve teachers of schools that experienced one of the events to understand which emotions students felt, thus creating a story and an integrated educational activity.
In order to cover press properly and to monitor what kind of articles are published and where, it is important to follow journalists during their visit.

A few years ago, a museum in the center of Italy closed for a while for the restoration of the building. A famous architect was involved in the project and an international curator was assigned with the launching exhibition for the reopening. The day before the public opening, press and cultural personalities were invited for a preview that also included a dinner, but the setup wasn't ready, the museum's areas were dirty, the dinner did not feature assigned seats and the whole night was a mess.

The worst outcome of this launching exhibition, one that people had been waiting for five years, was a terrible review published in one of the most important international art magazines that influenced a lot of colleagues from other media.

The largest number of information institutions collect, the better they know their audience and the more they can improve the cultural experience for the future.

References

1. Ariely, D.: twitter post in Jan 6th 2013
2. Churchill, W.: speech at The Royal Academy in London, April the 30th 1938
3. Gates, B.: Business @ the Speed of Thought: Succeeding in the Digital Economy, Grand Central Publishing (2009)
4. Musk, E.: https://www.forbes.com/sites/jonyoushaei/2018/03/06/elon-musks-10-secrets-to-success/?sh=57dde0e281e9. Cited in Mar 6, 2018 by Jon Youshaei on Forbes
5. Picasso, P.: quote generally attributed to
6. Warhol, A.: quote generally attributed to

Mobile Phone Data to Monitor the Impact of Social and Cultural Events of Brescia

Dati da telefonia mobile per monitorare l'impatto degli eventi sociali e culturali di Brescia

Maurizio Carpita, Marica Manisera, Paola Zuccolotto

Abstract In this short paper we describe our experience with mobile phone data to monitor the impact of social and cultural events. We show how information obtained with high time frequency from mobile phone networks can be very useful to detect people presence in urban centres, and gives support to manage citizen services, policies, and decision-making activities for *smart cities*.

Abstract *In questa nota descriviamo la nostra esperienza con i dati di telefonia mobile usati per il monitoraggio degli eventi sociali e culturali. Mostriamo come le informazioni da reti di cellulari sono molto utili per rilevare le presenze di persone nei centri urbani e offrire supporto alla gestione dei servizi ai cittadini, alle politiche e alle attività decisionali per le* smart cities.

Key words: crowding indicators, big data, Histogram of Oriented Gradients

1 Introduction

Nowadays smart technologies and big data have become a relevant part of the decision support systems for planning smart cities and offering smart services to citizens (see [3,4] and references cited therein). Mobile phone data can be used as a

[1] Maurizio Carpita, DMS StatLab, Dep. of Economics and Management (*dms-statlab.unibs.it*) and BODaI-Lab (*bodai.unibs.it*); email: maurizio.carpita@unibs.it

Marica Manisera, DMS StatLab, Dep. of Economics and Management (*dms-statlab.unibs.it*) and BODaI-Lab (*bodai.unibs.it*); email: marica.manisera@unibs.it

Paola Zuccolotto, DMS StatLab, Dep. of Economics and Management (*dms-statlab.unibs.it*) and BODaI-Lab (*bodai.unibs.it*); email: paola.zuccolotto@unibs.it

proxy of city users' presence, to assess how the city is crowded in different moments in time: a very useful information, especially in these COVID-19 times. This information is more useful than the number of residents provided by the standard census data, since it permits to produce dynamic instead of static crowding indicators and models. In this short paper, we summarize our experience with these types of big data, used with the aim to monitor social and cultural events in Brescia.

2 Data structure, data analysis methods and case studies

Our research with mobile phone signals started in 2013, thanks to the special agreement between the Municipality of Brescia and *Telecom Italia* (the biggest Italian telephone operator), which offered us the opportunity to analyse the presence of users connected to the Telecom Italia Mobile (TIM) network up to August 2016.

The first step was the acquisition and the organization of the row data. Fig. 1 represents the big data information flow related to all our analyses. The TIM mobile phone data *density* (estimated people presence for each time interval) has been detected on *pixels* (squares) with sides of 150 meters, each of which geo-referenced by the latitude and longitude of their *centroids* (Fig. 1, left). The technological platform of Telecom Italia has supported data *extraction*, *transformation*, *loading* (ETL) and *storage*, so that row data was correctly transferred to the server of the Municipality of Brescia (Fig. 1, centre). Finally, the DMS StatLab develops data analyses, maps, and reports useful for city's policy decisions (Fig. 1, right).

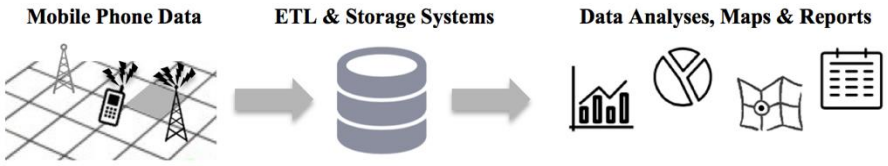


Figure 1: The mobile phone big data information flow representation

Our research with mobile phone data started in 2013; we have adopted a multi-stage approach to estimate the no. of daily phone users by means of the *Histogram of Oriented Gradients* (HOG) algorithm for data dimensionality reduction, and a mix of *k*-means and Functional Data Analysis (FDA) clustering methods for profiling time periods. HOG represents an image (a set of contiguous pixels or *raster*) as a unidimensional feature vector, that is analysed using standard statistical procedures. First, each raster is partitioned into sub-rasters, and for each pixel of them the *vector of gradients* $\boldsymbol{g} = (g_x, g_y)$ (differences right – left = g_x and up – down = g_y of densities around pixel) are computed. Second, for each $\boldsymbol{g}$, two measures are computed: *Direction* = $arctan(g_x / g_y)$ and *Magnitude* = $\| \boldsymbol{g} \| = (g_x^2 + g_y^2)^{1/2}$.

The final HOG object for each smaller raster is obtained binning the *Directions* and sum the *Magnitudes* for each bin; the final HOG vector of the full raster is obtained stacking the vectors of its smaller rasters. The matrix **X,** with the days of the period of interest on the columns and the stacked HOGs for the time intervals of each day on the rows, is created and used with the *k*-means and FDA cluster analysis to classify the daily profiles with respect to weekdays and months. We also developed a strategy to match mobile phone signals with census data (see [4,5] for details).

2.1 The Mille Miglia and the Giro d'Italia 2013

Our first research experience with mobile phone data has been the statistical analysis of TIM density in the city of Brescia during the week of the historic car race *Mille Miglia* (May 13-19, 2013) and the week that ended with the arrival of the national cycling race *Giro d'Italia* (May 20-26, 2013).

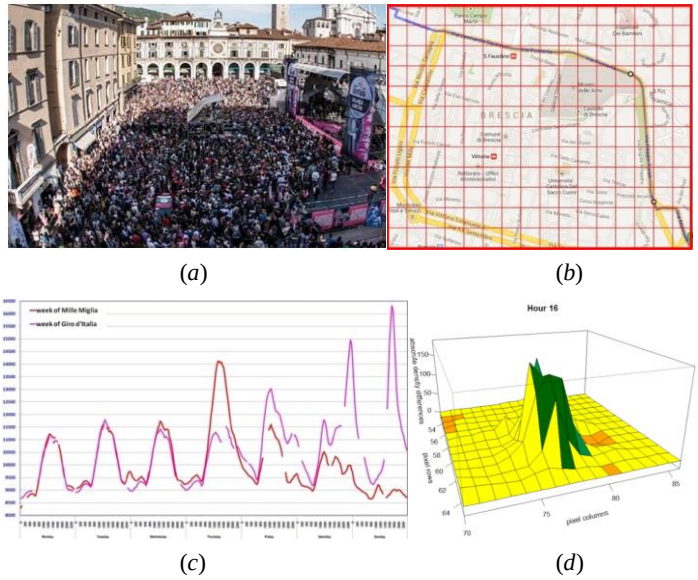


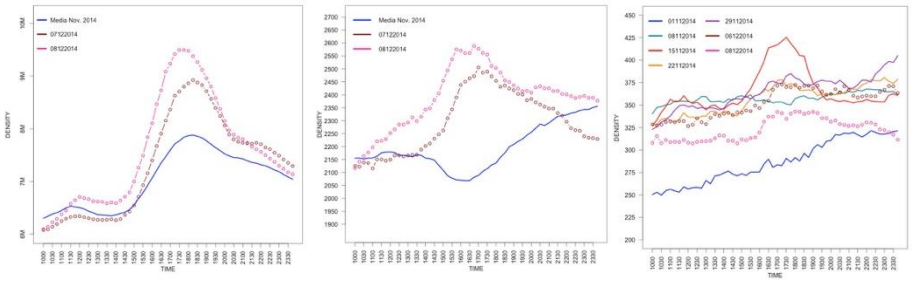
Figure 2: (*a*) People in *Piazza della Loggia* of Brescia during the award ceremony of the *Giro d'Italia* on May 26, 2013, (*b*) TIM pixels on the Brescia historic centre, (*c*) TIM density profiles in the week of the *Mille Miglia* and *Giro d'Italia* and (*d*) Differences at the day of the start of the *Mille Miglia* (May 16, 2013 at 4:00 PM) between TIM density of the day and benchmark (average of the other weekdays)

These were two social events of great impact, which attracted many people especially in the historic centre of the city: Fig. 2(*a*) shows the people in *Piazza della Loggia* of Brescia during the award ceremony of the *Giro d'Italia* on May 26,

2013. Our analysis was focused on an area of about 2.4 x 1.8 km as represented in Fig. 2(*b*), for which we have estimated the presence of people at intervals of 30 minutes (with some missing data): Fig. 2(*c*) shows TIM density profiles in the weekdays of the *Mille Miglia* and *Giro d'Italia*, while Fig. 2(*d*) shows differences at the day of the start of the *Mille Miglia* (May 16, 2013 at 4:00 PM) between the TIM density for the day of the event and for the benchmark (average of the other weekdays). For the *Mille Miglia*, the Municipality carried out three surveys aimed at tourists, business owners and residents: mobile phone data analysis, together with the results of these standard surveys, were used for planning the 2014 edition of this historic car race. In particular, the scarce attendance recorded on the evening of Saturday (the day of race arrival) have provided information support to the decision to move the awards ceremony on Sunday, to facilitate the organization of collateral events as the *Notte Bianca*. See [3] for more details.

2.2 The Christmas Markets 2014

In December 2014, TIM mobile phone data has been used to estimate the attendance of people in some places in the province of Brescia during the Christmas Market period. The goal was to compare the TIM density profile with time intervals of 15 minutes in the days of Christmas Market with a benchmark profile.



(*a*) *Brescia Centro* (*b*) *Carpenedolo* (*c*) *Darfo Boario Terme*

Figure 3: TIM density profiles of the Christmas Markets 2014 in (*a*) *Brescia Centro*, (*b*) *Carpenedolo* and (*c*) *Darfo Boario Terme*

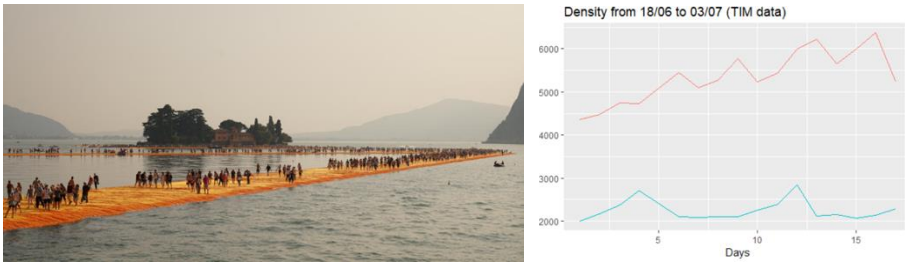
In Fig. 3(*a*), the TIM density profile for the Christmas Market of *Brescia Centro* (December 7 and 8, 2014) shows a significant increase in attendance for the two days of the event compared to the density average of Sundays in November 2014, the month considered as benchmark. In Fig. 3(*b*), the TIM density profile for the Christmas Market of *Carpenedolo* (December 7 and 8, 2014) shows a significant increase especially in the afternoon for the two days of the event compared to the density average of Sundays in November 2014. For the Christmas Market of *Darfo Boario Terme* (Sundays November 2014, from Saturday 6 to Monday 8 of December

2014), the TIM density profile in Fig. 3(*c*) clearly increases in the afternoon of Saturday 15 November 2014: after a short web search we found out that, in this day at 4:00 PM, the "*Festa Alpi Centrali*" took place (*fisialpicentrali.it/cms/sabato-15-novembre-2014-festa-alpi-centrali/*).

2.3 The Floating Piers 2016

From June 18 through July 3, 2016, the Italian Iseo Lake (about 100 kilometers east of Milan) was reimagined, with an international event free and open to people: a temporary art installation, a 3-kilometer-long walkway on the water named *The Floating Piers*, was created using canvases, cables, and metal structures by the contemporary artist Christo Vladimiroff Javacheff (1935-2020); see Fig. 4(*a*). Local authorities estimated that 1,2 million people visited the site in the sixteen days of the event, an average of 72,000 visitors daily in an area where there usually are about 12,000 residents. Other sources estimated 1,5 million visitors, with a daily average of 100,000 attendances and a peak of 115,000 attendances reached on July 1.

In [2] the TIM densities for intervals of 15 minutes on three areas of the Iseo Lake (Monte Isola, Sulzano and Iseo) have been used to estimate the *Benchmark profile* (data for days of June and July 2014 and 2015) and the *Floating Piers profile* (data from June 16 to July 3, 2016) of people presences.



(*a*) (*b*)

Figure 4: (*a*) *The Floating Piers* (*christojeanneclaude.net/projects/the-floating-piers*) and (*b*) *Floating Piers* (higher) and *Benchmark* (lower) TIM density profiles for the 16 days of the event

Fig. 4(*b*) shows the impact of Christo's The Floating Piers on the Iseo Lake during the 16 days of the event: *The Floating Piers profile* is from 2 to 3 times the *Benchmark profile* and the number of people that visited the piers on the water increase along the period. We applied the TIM multiplicative factor of five to these two *profiles* [5,6], in the area of the Iseo Lake: for the second half of June, the estimated benchmark range is 10,000-15,000 people daily, whereas from June 18 through July 3 in 2016, the estimate of the daily attendances increases to 22,500-32,500 (+45%). Since the installation was open from 8 am to 10 pm (14 hours),

assuming a visit's average time of about 4 hours, the median of the daily number of visitors of *The Floating Piers* is estimated from about 78,000 to about 115,000 on Friday July 1, 2016, results consistent with the official statistics.

3 Actual projects and future research

Our actual project with mobile phone data, named DS4BS (*Data Science for Brescia*) and started in September 2021, has the main objective to increase knowledge about the way people visit the cultural places (museums, theaters, monuments, and historic buildings), to support institutions and decision makers. Special attention is devoted to the experimentation of new ways for public detection and engagement, exploring cultural attitudes and perceptions, and developing new forms of accessibility to culture, also with reference to the cultural and sustainable tourism. A research line is devoted to test innovative methods of integration and analysis of big data from mobile phone and social network systems, in order to collect useful information for monitoring visits to some cultural sites of Brescia. The DS4BS Project has received the endorsement of the *Municipality of Brescia* and the *Fondazione Brescia Musei* and is part of the research activities dedicated by the University of Brescia to support the event planned for 2023, when Bergamo and Brescia will be the *Italian Capital of Culture*. More information is available at: *dms-statlab.unibs.it/ricerca-e-pubblicazioni/progetto-mub* and *bodai.unibs.it/ds4bs/*.

Data from mobile phones can be used not only to estimate people *presences* in social and cultural events, but also people *flows*: for the MoSoRe (*Infrastrutture e servizi per la mobilità a sostenibile e resiliente*) Project, co-founded by *Regione Lombardia* (*bit.ly/2Xh2Nfr*), we used origin-destination data from TIM mobile phones provided by *Olivetti* (*www.olivetti.com/en/iot-big-data*) and *FasterNet* (*www.fasternet.it*) on hourly basis for twelve months (from September 2020 to August 2021) to analyse the *Mandolossa* area (a critical zone with flood episodes in the north-west of the city of Brescia), with the aim to predict the amount of traffic flows in the context of emergency management plans [1].

Acknowledgements This research at DMS StatLab and BODaI Lab of the University of Brescia is co-funded for the DS4BS Project (Rif. 2020-4334) by Fondazione Cariplo and for the MoSoRe Project by Regione Lombardia, Italy (CallHub ID 1180965).

References

Mobile Phone Data to Monitor the Impact of Social and Cultural Events of Brescia

1. Balistrocchi, M., Metulini, R., Carpita, M., Ranzi R.: Dynamic maps of people exposure to floods based on mobile phone data. Nat. Hazards and Earth Syst. Sciences **20**, 3485-3500 (2020)
2. Carpita, M. (2019) The mobile phone big data tell the story of Christo's The Floating Piers impact on the Lake Iseo. In: Carpita, M., Fabbris, L. (eds.) Book of Short Papers of the ASA Conference on Statistics for Health and Well-being, pp. 53-56. CLEUP, Padova (2019)
3. Carpita, M., Simonetto, A.: Big data to monitor big social events: analysing the mobile phone signals in the Brescia smart city. Electronic J. of Appl. Statistical Analysis: DSS **5**(1), 31-41 (2014)
4. Metulini, R., Carpita, M.: A Spatio-temporal indicator for city users based on mobile phone signals and administrative data. Soc. Indicators Res. **156**(2-3), 761-781 (2021)
5. Metulini, R., Carpita, M.: A strategy for the matching of mobile phone signals with census data. In: Arbia G., Peluso S., Pini A., Rivellini G. (eds.) SIS 2019 Smart Statistics for Smart Applications Book of Short Papers, pp. 427-434. Pearson Publ., Milano (2019)

Analysing preferences on sustainable tourism in Vallo di Diano area, Campania, Italy

Una analisi delle opinioni sul turismo sostenibile nel Vallo di Diano, Campania, Italia

Stefania Capecchi, Giovanni Quaranta and Rosanna Salvia

Abstract The paper investigates the response patterns of preferences towards sustainable tourism in Vallo di Diano area, Campania region, southern Italy, collected on a sample of local actors. Exploiting a model-based framework, results show that both *contextual* and *subjective* characteristics significantly affect the expressed preferences.

Abstract *Il lavoro analizza le preferenze espresse da un campione di rispondenti del Vallo di Diano, in Campania, Italia, rispetto al tema del turismo sostenibile. Mediante un approccio modellistico, emerge che sia le variabili contestuali che quelle individuali influenzano in misura significativa le preferenze espresse.*

Key words: Sustainable tourism, Preferences data, CUB models

1 Introduction

Tourism has been one of the fastest growing industries in the world in the past 50 years, and although it generates considerable economic effects, there is a prominent concern that it may function as a double-edged sword. This is more evident in case of protected areas: large numbers of visitors may positively influence the development of territories, although they could also dangerously impact on natural

Stefania Capecchi
Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò 22, I-80138 Naples, Italy e-mail: stefania.capecchi@unina.it

Giovanni Quaranta
Department of Mathematics, Informatics and Economics, University of Basilicata, Macchia Romana, I-85100 Potenza, Italy e-mail: giovanni.quaranta@unibas.it

Rosanna Salvia
Department of Mathematics, Informatics and Economics, University of Basilicata, Macchia Romana, I-85100 Potenza, Italy e-mail: rosanna.salvia@unibas.it

and cultural heritage of destinations. The symbiotic relationship between tourism and land use may also lead to prominent changes in territory exploitation, such as soil and environmental degradation, water pollution, and biological diversity loss [2]. Sustainable tourism has become an established area of academic interest and it has been adopted into tourism policy-making by public and private sectors at all levels of governance [7]. Moreover, plenty of options are offered to policymakers to encourage tourism while protecting and enhancing the future opportunities of both visitors and host countries. In Italy, statistics on the pressure on the environment of tourism activities are produced by ISTAT [3], also due to the Sustainable Development Goals initiative of the United Nations [6].

This study provides an analysis of subjective evaluations expressed by a large number of respondents, mostly local actors from the rural and mountain district of Vallo di Diano, partly overlapping with the National Park of Cilento and Vallo di Diano territory, a well-known tourist venue located in the Province of Salerno, in southern Italy. Relying on a class of mixture models, available information and selected methodology are illustrated; then the empirical results are discussed. Some concluding remarks end the study.

2 Data

Promoted and managed by an integrated *Local Action Group* (GAL), named "GAL Vallo di Diano" (a consortium of municipalities, public and private local stakeholders), an observational study was carried out in 2016. Main goal of the survey was to collect local actors' opinions on several major issues raised by the development prospects of the territory. An on-line questionnaire was administered to respondents from 27 municipalities of the Vallo di Diano area, which is mostly a rural district. Respondents were asked to provide socio-demographic (*subjective*) information and to express, on a 7-point rating scale (1=lowest, 7=highest preference), their preference level about 11 thematic fields/items: 1) Development/innovation of supply chains and local production systems; 2) Development of renewable energy; 3) Sustainable tourism; 4) Care and protection of landscape, land use and biodiversity; 5) Enhancement of cultural and artistic heritage; 6) Access to essential public services; 7) Management of environmental and natural resources; 8) Social inclusion of disadvantaged and/or marginal groups; 9) Legality; 10) Urban re-development; 11) Smart networks and communities.

In addition, official statistics on (*contextual*) characteristics were collected for each municipality.

After a preliminary screening, dataset consists of $1,383$ observations. Since the questionnaire was administered on-line, a selection bias occurs; therefore, a comparison was made between the characteristics of the respondents and those available from the public archives of the respective municipality.

Thus, women represent 42.7% of the sample and they are younger than men. Marital status is comparable by gender, while women are definitely more educated

than men: 34% of women hold a tertiary degree against 25% of men, who are instead more frequently employed. Only 21% of respondents are farm owners (this percentage decreases to 14% for women). Participation in associations and groups is frequently reported (41%), nonetheless about $1/3$ of the sample declares no knowledge of GAL, whereas 58% of respondents (50% for women) positively evaluate its existence and activities. About $2/3$ of interviewees declare to experience an overall good economic situation (on a 10-point scale), a variable employed as a proxy of respondents' income.

As far as contextual variables are concerned, the population density of municipalities ranges from 12 up to 493 inhabitants/km^2, and active population rate is in between 52% to 74%, with a limited percentage of workers in agricultural activities. The rate of those employed in services varies from less than 1% up to 8%. The percentage of agriculture production areas is generally high and a number of farms and zoo-technical firms are present in several municipalities, while 9 out of 27 municipalities may be classified as *Mountain* municipalities. Some development initiatives financed by GAL and by the regional Rural Development Program (RDP) are present in the whole district. Overall, secondary education at the municipality level ranges from 21% up to 32%, whereas tertiary education rate is generally low, varying from 3% to 12%.

Thus, compared to the indices registered for each municipality, the sample consists of respondents more educated and wealthy than the average.

The preferences for the 11 thematic fields reach high and very high ratings, especially for items 1, 3 and 11; in all the circumstances, a low variability turned out. Then, a Cronbach $\hat{\alpha} = 0.886$ may be interpreted as a very good internal consistency measure.

In this exercise, we focus on item 3, since the issue of sustainable tourism has been largely discussed in the Vallo Diano district, whose territory is partly overlapping with a protected area, and it has also been widely debated in literature [1].

3 Selected models

A model-based approach may be implemented to investigate response patterns of preferences, that is the process by which an interviewee selects a categorical choice out of a list of ordered descriptions, and to detect possible drivers for the selected category. The specific mixture we selected accounts for both the expressed consensus towards the item (denoted as *feeling*) and the inherent indecision which always accompanies any selection process (denoted as *uncertainty*).

Formally, for a given number $m > 3$ of categories, a CUB model is a *C*ombination of a shifted *B*inomial and a discrete *U*niform distributions, respectively, firstly introduced by [4] and fully discussed and generalized in [5]. Then, the probability of selecting the r-th category for the i-th subject is:

$$Pr(R_i = r|\, y_i, w_i) = \pi_i \left[\binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} \right] + (1-\pi_i) \left[\frac{1}{m} \right],$$

for $r = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$. The (common and suitable) logistic links between the parameters $\pi_i \in (0,1)$, $\xi_i \in (0,1)$ and the subjects' covariates are:

$$\begin{cases} \pi_i = \dfrac{1}{1+e^{-y_i\beta}} \\ \xi_i = \dfrac{1}{1+e^{-w_i\gamma}} \end{cases} \Longleftrightarrow \begin{cases} logit(1-\pi_i) = -y_i\beta = -\beta_0 - \beta_1 y_{i1} - \cdots - \beta_p y_{ip}; \\ logit(1-\xi_i) = -w_i\gamma = -\gamma_0 - \gamma_1 w_{i1} - \cdots - \gamma_q w_{iq}; \end{cases}$$

where y_i and w_i are the row vectors containing the covariates values of the i-th subject, for $i = 1, 2, \ldots, n$.

In this way, each interviewee responds with different weights π_i and $1-\pi_i$ to *feeling* and *uncertainty* components. In fact, CUB models are more parsimonious than the classical ones (since they do not require the estimation of cut-points) and can be checked for the significance of subjective and contextual covariates as drivers of uncertainty and feeling, respectively. Finally, given the one-to-one correspondence between the whole probability distribution and the parameters $(1-\pi, 1-\xi)$ defined over the unit square, a remarkable advantage of these models is their graphical visualization tools.

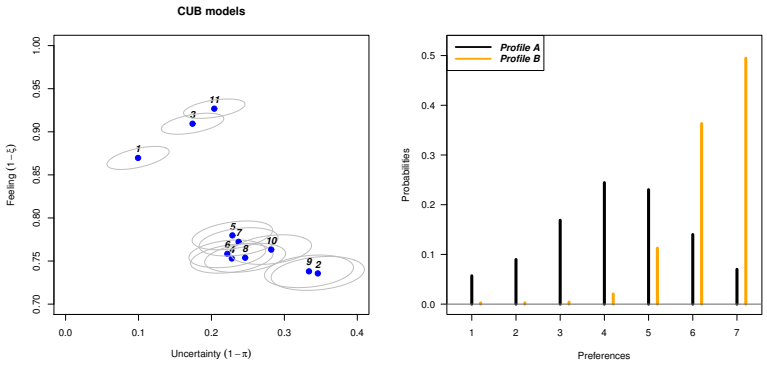
4 Results

To compare the sustainable tourism item with the others, panel (a) of Figure 1 represents estimated CUB models for the 11 thematic fields (without covariates). Confidence ellipses clearly overlap in several cases and show few clusters among the response patterns: for instance, higher *feeling* and low uncertainty towards 1, 3 and 11 items, versus greater *uncertainty* and lower *feeling* towards all the others.

Next, focusing on item 3 (Sustainable Tourism), several CUB models have been estimated and subjective and/or contextual variables are selected by stepwise procedures based on significance and likelihood criteria. For lack of space, only the best solution is reported with estimated parameters (standard errors in parentheses) as obtained by package `CUB` in *R* environment.

$$\begin{cases} logit\,(1-\hat{\pi}_i) &= \underset{(0.760)}{-1.720} + \underset{(5.512)}{22.463}\, Services + \underset{(0.399)}{1.357}\, Dfarm_i + \\ & \quad \underset{(0.113)}{-0.319}\, Econ_i + \underset{(0.413)}{1.480}\, Assoc_i - \underset{(0.431)}{1.598}\, Useful_i; \\ logit\left(1-\hat{\xi}_i\right) &= \underset{(0.494)}{-0.221} + \underset{(0.001)}{0.003}\, Popdens - \underset{(6.330)}{16.315}\, Agrw - \underset{(2.137)}{17.040}\, Services - \underset{(0.011)}{0.099}\, RDP + \\ & \quad + \underset{(2.102)}{11.372}\, SecEduc - \underset{(2.605)}{9.526}\, Univ + \underset{(0.684)}{3.616}\, Gender_i + \underset{(0.029)}{0.267}\, Econ_i + \\ & \quad + \underset{(0.117)}{0.559}\, Assoc_i + \underset{(0.095)}{0.260}\, Useful_i - \underset{(0.181)}{0.932}\, GAge_i; \end{cases}$$

Preferences on sustainable tourism



(a) (b)

Fig. 1 (a) Estimated CUB models of the 11 thematic fields; (b) Two typical respondent profiles

The *uncertainty* link shows that running a farm (`Dfarm`) and participating in associations and groups (`Assoc`) are positive drivers. Also, uncertainty is lower for people who evaluate the GAL activities as (`Useful`) and for increasing level of the income proxy (`Econ`). `Services`, the only significant contextual covariate, exerts a positive effect on the inherent indecision.

The *feeling* link shows a positive (increasing) impact of women (in case `Gender`=1), reported economic situation (`Econ`), participation in associations and groups (`Assoc`), and positive consideration of GAL activities (`Useful`). The interaction between gender and age (`GAge`) indicates that older women express a lower preference towards this item. Among the contextual variables, an increasing (positive) effect derives from population density (`Popdens`) and secondary education rate (`SecEduc`) at municipality level, while a decreasing effect is associated to tertiary education rate (`Univ`) as well as to the number of development initiatives (`RDP`), the rate of people employed in `Services` and in Agriculture (`Agrw`).

To visualize the impact on the expected profile of responses for two different subjects living in different contexts, panel (b) of Figure 1 reports the probabilities of preferences for two quite diverse persons. *Profile A* refers to a 60-year-old male, experiencing a quite precarious economic situation, not participating in groups and associations and not considering useful the GAL activities, who lives in a municipality where the initiatives of the regional rural development plan are scarce and the percentage of service employees is 0.08%. *Profile B* represents a 30-year-old woman, with an excellent economic situation, very active in associations and happy with GAL activities, resident in a municipality in which the initiatives of the regional rural development plan are numerous, and services are 0.03%. The remaining variables are set to modal or average values for both profiles.

As a consequence of the estimated model, the distribution of profile A is almost symmetrical with modal value at 4, thus expressing a substantial indifference

towards sustainable tourism. On the contrary, profile B shows a probability concentration on higher ratings, with a modal value at 7.

5 Concluding remarks

Model-based results show that both contextual and subjective features are explanatory of the preference patterns on sustainable tourism. This circumstance implies that geographical, cultural and economic aspects are relevant as well as the subjective ones. The contextual aspects are of course more difficult to modify in the short term, but their significance may stimulate stakeholders and policy-makers to address possible local development interventions. In this respect, two distinct respondents' profiles emphasize how subjective and contextual diversities are jointly responsible for quite different responses. Such findings seem to confirm that the problems related to the tourism development of the territory are more strongly perceived by young respondents, at the beginning of their working life, and mostly by women. Further investigation is needed to generalize such an observational study to capture more representative samples of the whole population of these areas and provide more explicative power.

References

1. D'Arco, M., Lo Presti, L., Marino, V., Maggiore G.: Is sustainable tourism a goal that came true? The Italian experience of the Cilento and Vallo di Diano National Park. Land Use Policy **101**, 1–12 (2021)
2. Heslinga, J.H., Groote, P., Vanclay, F.: Using a social-ecological systems perspective to understand tourism and landscape interactions in coastal areas. J. Tour. Fut. **3**, 23–38 (2017)
3. ISTAT: Conti integrati economici ed ambientali del turismo: pressioni delle attività turistiche sull'ambiente naturale. Statistiche Sperimentali (2019). Available at http://www.istat.it/it/files//2019/03/principali-risultati-e-nota-metodologica.pdf
4. Piccolo, D.: On the Moments of a Mixture of Uniform and Shifted Binomial random variables. Quad.Stat. **5**, 85–104 (2003)
5. Piccolo, D., Simone, R.: The class of CUB models: statistical foundations, inferential issues and empirical evidence, with dicussions and rejoinder. Stat. Methods Appl. **28**, 389–493 (2019)
6. United Nations: Division for sustainable development goals (2021). Available at http://www.sdgs.un.org/topics/sustainable-tourism
7. Zolfani, S.H, Sedaghat, M., Maknoon, R., Zavadskas, E.K.: Sustainable tourism: a comprehensive literature review on frameworks and applications. Economic Research-Ekonomska Istraživanja **28**, 1–30 (2015)

Session of free contributes SCL10– *Time Series data, Panel data and Circular Economy*
Chair: Anna Crisci

Testing the Participation Gap Inclusion within the Wage Phillips Curve

Effetti dell'inserimento del gap nella partecipazione all'interno della curva di Phillips

Deborah Scaccabarozzi, Daniele Toninelli, Davide Zurlo
Fabio Bacchini, Roberto Iannaccone

Abstract The Phillips curve alternative formulations have been widely studied, over the last years. Our research analyzes the Phillips curve in the framework of the MeMo-It model, used by Istat to provide forecasts for the Italian economy. In particular, we study an alternative specification including a measure of the labor market slack, i.e. the participation gap. This variable is computed as the difference between trend and actual labor force. The obtained specification of the Phillips curve allows analyzing how an improvement in the participation gap can lead to positive effects on wage growth. Our findings about the new formulation report a slight increase in the goodness of fit and an enhanced forecasting performance, in terms of prediction accuracy.

Abstract *Diverse formulazioni della curva di Phillips sono state oggetto di studio, nel corso degli ultimi anni. Il nostro lavoro analizza la curva di Phillips nello specifico contesto del modello MeMo-It, utilizzato da Istat per fornire previsioni per l'economia italiana. In particolare, studiamo una specificazione alternativa che include una misura aggiuntiva del livello di rallentamento nel mercato del lavoro: il gap nella partecipazione. Questa variabile è calcolata come la differenza tra il trend della forza lavoro e la forza di lavoro effettiva. Questa specificazione della curva di Phillips consente di analizzare come un miglioramento nel gap della partecipazione possa condurre a degli effetti positivi sulla crescita del salario. I nostri risultati mostrano un lieve incremento nella bontà delle stime ed una migliore capacità di previsione, in termini di accuratezza.*

[1] University of Bergamo, deborascaccabarozzi96@gmail.com; daniele.toninelli@unibg.it
Istat, zurlo@istat.it; bacchini@istat.it; iannacco@istat.it

Key words: wage growth, labor market slack, Phillips curve, participation gap.

1 Introduction

In the original formulation of the Phillips curve (Phillips, 1958) it was identified a negative relationship between wage inflation and unemployment, for the UK. Later specifications include the expectations-augmented Phillips curve (Friedman, 1968), the New Classical Phillips curve (derived by Lucas' surprise aggregate supply function; see Lucas, 1973), the New Keynesian Phillips curve (Roberts, 1995), the Hybrid Phillips curve (Galì and Gertler, 1999).

In this work we aim at extending the existing literature by presenting an analysis of the Phillips curve in the framework of the MeMo-It model. MeMo-It is the macroeconometric model developed by Istat (the Italian National Institute of Statistics) to provide forecasts for the Italian economy. MeMo-It is based on a mixture of the London School of Economics and of the Fair-updated Cowles Commission approaches. In order to merge theory and data, MeMo-It uses cointegration methods on dynamic sub-systems to estimate theory-interpretable and identified steady state relationships, imposed in the form of equilibrium-correction models (see Bacchini *et al*., 2013, Bacchini *et al*., 2018).

In our research, we use Italian annual data provided by Istat for the period 1980 to 2019. The new curve specification tested here includes an additional variable: the participation gap. With this new specification we can assess how an improvement in this last variable is able to affect wage growth. Therefore, this is a useful method to evaluate the labor market status.

We assess the validity of the new Phillips curve equation by means of the goodness of fit and in terms of its forecasting performance. More into details, we compare the results of this work and the ones obtained in Scaccabarozzi *et al*. (2021), in order to find the best-performing alternative Phillips curve specification.

Section 2 of this paper introduces the literature review as well as our methodology. Section 3 presents the main results of our study. Section 4 proposes our conclusions.

2 Literature Review and Methodology

The Phillips curve has been studied according to two main strands: at the macroeconomic level, by performing time series analysis (e.g., Bhattarai, 2016, and Bulligan et al., 2017) or at microeconomic level, by performing panel data analysis (Bjørnstad and Nymoen, 2008). Our work follows specifically the first approach, focusing on a macroeconomic perspective.

Testing the Participation Gap Inclusion within the Wage Phillips Curve

Several works addressed the effective existence of the Phillips curve (e.g., Bhattarai, 2016) or extended the Phillips curve equation by considering the role of some additional factors (e.g., Conti and Gigante, 2018). A recent issue emerged in the literature: how to measure the economic slack. This is a relevant issue, because if a portion of slack is not captured by the traditionally used measures (unemployment and output gap), the resulting estimates of the Phillips curve parameters will be affected. Recently it was found that unemployment could provide only a partial picture of the real labor market situation (Eurofound, 2017). Indeed, among employed people, there is the category of involuntary part-time workers. Then, among inactive people, there are people who are not seeking work, but are available, and people who are seeking work, but are not immediately available. Istat provides these measures starting from 2004. Originating from these considerations, other measures of labor slack were conceived. Bulligan *et al.* (2017), for example, measured the labor slack through the EC indicator of labor shortage and hours per worker. Using U.K. data, Bell and Blanchflower (2014) found that long-term unemployment does not have any effect on wage determination. Blanchflower and Levin (2015) introduced another measure: the employment gap, estimated as the sum of the unemployment gap, the underemployment gap and the participation gap. The latter is the difference between actual and potential labor force. The analysis of the participation rate was highlighted also by Fay and Ketcheson (2016). Nickel et al. (2019) include the UCM participation rate gap among the unconventional measures of labor market stance. Smith (2014) analyzed different measures of labor slack, because some individuals who do not participate to the labor market have more possibility to enter again, if compared to other non-participants.

In our work, we have to deal with an issue about data availability: data regarding the specific labor slack categories were only recently made available, and they do not cover our whole period of analysis (1980 to 2019). We focus on participation gap, calculating it as the difference between trend labor force, estimated using the Hodrick-Prescott filter (Nilsson and Gyomai, 2011), and actual labor force. The presence of labor slack occurs when actual labor force is lower than its trend. This is a way to measure a part of the labor slack categories. If labor force is lower than trend, it means that there is a proportion of people that could be included in the labor force, that currently is not participating in the labor market. In particular, such people could include the ones looking for a job but that would be available to work if the conditions were favorable.

The original specification expresses the differenced log-transformation of wages (*WIPCP*) as function of prices (*PCH*), unemployment gap (UR_t-*NAWRU*), productivity (YO_{t-1}/ULA_{t-1}) and a dummy variable for years 2010 and 2015 (*D2010+d2015*). The explanatory variables are log-transformed and first-differenced, with the exception of unemployment gap. The alternative equation is expressed as follows:

$$\Delta \log(WIPCP) = \beta_1 \Delta \log(PCH_{t-1}) + \beta_2 \Delta \log\left(\frac{YO_{t-1}}{ULA_{t-1}}\right) + \beta_3 * \left(\frac{UR_t - NAWRU}{100}\right) + \beta_4 (D2010 + D2015) + \beta_5 PART\_GAP,$$

where *PART_GAP* is the difference between trend and actual labor force.
The methodology followed in this work is aligned to the one adopted by Scaccabarozzi *et al.* (2021). The models, estimated using the Ordinary Least Squares (OLS), are compared in terms of goodness of fit (i.e., using the Adjusted R^2) and in terms of forecasting performance. The latter is assessed by four measures: the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE). The results are also compared to the ones presented in Scaccabarozzi *et al.* (2021), in order to determine which Phillips curve shows a better performance.

3 Results

Table 1 in *Column 1* reports the results from the OLS estimation of the current MeMo-It Phillips curve equation (*Mod1*); *Column 2* (*Mod2*) shows the alternative specification results, including the participation gap.

Table 1: OLS estimates for the original MeMo-It Phillips curve (*Mod1*) and for the alternative specification with the inclusion of the participation gap (*Mod2*)

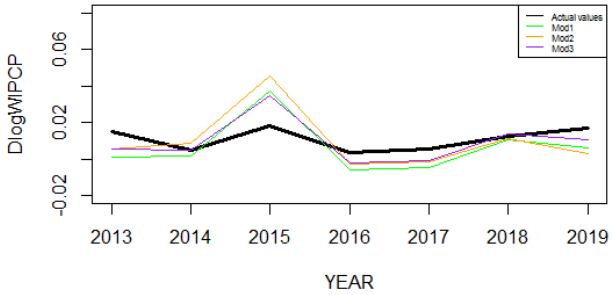
Coefficients	***Mod1***	***Mod2***
Lag(DlogPCH)	0.875***	0.872***
	(0.040)	(0.039)
UR_T_NAWRU	-0.003**	-0.003**
	(0.002)	(0.002)
PART_GAP		-0.590*
		(0.310)
dummy	0.034***	0.037***
	(0.158)	(0.152)
Observations	38	38
R^2	0.961	0.965
Adjusted R^2	0.956	0.959
Residual Std. Error	0.012 (df = 34)	0.011 (df = 33)
F Statistic	207.414*** (df = 4; 34)	179.482*** (df = 5; 33)

Note: *p<0.1; **p<0.05; ***p<0.01

All coefficients have signs aligned with the literature. The inflation and productivity growth coefficients have a positive sign, showing a positive relationship between these variables and wage growth. The opposite situation is observed for the

coefficient of unemployment gap: it shows a negative sign. This is also consistent with expectations: a higher participation gap corresponds to a labor force lower than trend, therefore the labor market slack increases, leading to a reduction in wage growth. Most of coefficients are significantly different from zero at the 5% significance level, whereas the participation gap is significantly different from zero only at a significance level of 10%. The goodness of fit for the new model is similar to one reached by the specification presented in Scaccabarozzi *et al.* (2021), where the equation is extended with the inclusion of an error correction term. The values of the adjusted R^2 (both around 0.959) show a similar performance of the two alternative specifications. In both cases, if compared to the original model, there is a slight increase for this index (+0.31%). By means of MAE, MAPE, MSE and RMSE we compare the actual values with the predicted values. The latter are computed by means of the forward chaining method, a type of cross-validation that allows to account for the temporal dependencies present in the data, seen that we are dealing with time series. The test observations were computed for the years 2013 to 2019. *Figure 1* reports the actual values and the fitted values for the original specification (*Mod1*), for the specification with the inclusion of the participation gap (*Mod2*) and for the specification including the error correction term (*Mod3*).

Figure 1: Actual and predicted values for *Mod1*, *Mod2* and *Mod3*



In general, values predicted by *Mod3* are closer to the actual ones, if compared to both *Mod1* and *Mod2*. This conclusion is further confirmed by the following values obtained for the various measures of prediction accuracy:

- MAE: *Mod1*: 0.00993, *Mod2*: 0.01004 (+1.19%);
- MAPE: *Mod1*: 1.12964 *Mod2*: 0.98037 (-13.21%);
- MSE: *Mod1*: 0.00013 *Mod2*: 0.00016 (+29.25%);
- RMSE: *Mod1*: 0.01129 *Mod2*: 0.01283 (+13.69%).

If we include the participation gap, MAPE is even lower than the one observed for the original model. Contrarily, the other indexes show higher values. We highlight also that the four indexes computed for the specification including the error correction term (Scaccabarozzi *et al.*, 2021) are all lower than the ones observed for the specification with the participation gap and for the original specification.

4 Conclusions

Our research goal consists in evaluating an alternative reliable specification for the Phillips curve equation present in the Istat MeMo-It model. In particular, we aim at improving the currently used specification goodness of fit and its forecasting performance. Our alternative specification includes an additional measure of labor market slack, i.e., the participation gap, introduced in order to assess the impact of this additional variable on wage growth. The latter is a fundamental variable for labor market assessment and it is directly linked to citizens' well-being. Our specification reaches a goodness of fit similar to the one obtained using the specification tested by Scaccabarozzi *et al.* (2021). All coefficients, including the one of the additional variable, are significantly different from zero. Nevertheless, the forecasting ability of the new specification outperforms the original one only in terms of MAPE, whereas, according to the other three indexes, we obtain better results with the original model. If we take into account the four indexes assessing the forecasting ability, we should prefer the specification including the error correction term (Scaccabarozzi *et al.*, 2021), that is outperforming the two alternatives. We suggest analyzing a larger time span, including new data, in order to obtain further insights about these and others alternative Phillips curve specifications.

References

Bacchini, F., Bontempi, M. E., Golinelli, R., & Jona-Lasinio, C.: Short-and long-run heterogeneous investment dynamics. Empirical Economics, **54** (2), 343-378 (2018)

Bacchini, F., et al.: Building the core of the Istat system of models for forecasting the Italian economy: MeMo-It. Rivista di Statistica Ufficiale N. 1/2013 (2013)

Bell, D., & Blanchflower, D.: Labour Market Slack in the UK. National Institute Economic Review, **229**, F4-F11 (2014)

Bhattarai, K.: Unemployment–inflation trade-offs in OECD countries. *Economic Modelling*, **58**, 93-103 (2016)

Bjørnstad R. & Nymoen R.: The New Keynesian Phillips Curve Tested on OECD Panel Data. Economics: The Open-Access, Open-Assessment E-Journal, **2** (2008-23), 1–18 (2008)

Blanchflower, D. G., & Levin, A. T.: Labor market slack and monetary policy. Working Paper No 21094, National Bureau of Economic Research (2015)

Bulligan, G., & Viviano, E.: Has the wage Phillips curve changed in the euro area? IZA Journal of Labor Policy, **6** (1), 1-22 (2017)

Conti, A. M., & Gigante, C.: Weakness in Italy's core Inflation and the Phillips curve: The role of labour and financial market indicators. Bank of Italy Occasional Paper, No 466 (2018)

Eurofound: Estimating labour market slack in the European Union. Publications Office of the European Union, Luxembourg (2017)

Fay, R., & Ketcheson, J.: The US Labour Market: How Much Slack Remains? Staff Analytical Note 2016-9, Bank of Canada (2016)

Friedman, M.: The Role of Monetary Policy. American Economic Review, **58** (1), 1-17 (1968)

Galì, J., & Gertler, M.: Inflation dynamics: A structural econometric analysis. Journal of monetary economics, **44** (2), 195-222 (1999)

Lucas, R. E.: Some International Evidence on Output-Inflation Tradeoffs. The American Economic Review, **63** (3), 326–334 (1973)

Testing the Participation Gap Inclusion within the Wage Phillips Curve

Nickel, C., Bobeica, E., Koester, G. B., Lis, E. & Porqueddu, M.: Understanding Low Wage Growth in the Euro Area and European Countries. ECB Occasional Paper No. 232 (2019)

Nilsson, R. & Gyomai, G.: Cycle Extraction: A Comparison of the Phase-Average Trend Method, the Hodrick-Prescott and Christiano-Fitzgerald Filters. OECD Statistics Working Papers, OECD Publishing, No 2011/4 (2011)

Phillips, A. W.: The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957. Economica, **25** (100), 283-299 (1958)

Roberts, J. M.: New Keynesian Economics and the Phillips Curve. Journal of money, credit and banking, **27** (4), 975-984 (1995)

Scaccabarozzi, D., Toninelli, D., Zurlo, D., Bacchini, F., & Iannaccone, R.: Testing New Versions of the Wage Phillips Curve in the MeMo-It Model Used by Istat. In JSM Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association (2021)

Smith, C. L.: The Effect of Labor Slack on Wages: Evidence from State-Level Relationships. FEDS Notes 2014-06-02-2, Board of Governors of the Federal Reserve System (U.S.) (2014)

Multisource approach for trends evaluation. An application at the agricultural sector.

Approccio multi-fonte per la valutazione dei trend. Una applicazione al settore agricolo.

Daniela Fusco and Maria Antonietta Liguori and Valerio Moretti

Abstract The need to develop a system of homogeneous and comparable statistics, combined with the need to have updated statistics, has carried the adoption of statistical registers in Statistical Research Institutes. The dimension of information available is comparable with census data, giving the opportunity to follow the units observed over time and to evaluate in detail the evolution of the various sectors. The work applies this approach to Italian farms, defining their resilience in economic terms and production specialization. In particular, the analysis will focus on a panel of about 500,000 farms, which from 2000 to 2017 did not change their corporate composition.

Abstract *L'esigenza di sviluppare un sistema di statistiche omogene e confrontabili nel tempo unita alla necessità di disporre di statistiche aggiornate ha portato all'adozione dei registri statistici negli Istituti di Ricerca. Un patrimonio informativo di così ampia portata può essere confrontato con i dati censuari, dando la possibilità di seguire le unità osservate nel tempo e valutare dettagliatamente l'evoluzione dei diversi settori. Il lavoro applica questo approccio alle aziende agricole italiane, definendone la resilienza in termini economici e di specializzazione produttiva. In particolare l'analisi si concentrera su un panel di circa 500 mila aziende, che dal 2000 al 2017 non hanno modificato la loro composizione societaria.*

Key words: Record linkage, Data analysis, Agriculture

[1] Daniela Fusco, Istat; email: dafusco@istat.it
Maria A. Liguori, Istat; email: liguori@istat.it
Valerio Moretti, Istat; email: vmoretti@istat.it

1 Introduction

The modernisation process at many national statistical institutes (NSIs) whereby the goal is to make best use of available data, reduce the response burden and production costs [1]. The result is the increasing of common use administrative data sources to produce statistics.

For the agricultural sector, the most comprehensive source of the structure of Italian farms is the Census. In this work, the data from the V and VI General Census of Agriculture are be taken into account and compared together with the data from the Statistical Register of Agricultural Farms (Farm Register), in order to describe the changes in Italian agriculture from 2000 to today.

The attention was focused on a panel of about 500,000 farms distributed throughout the country that have maintained the same corporate structure since 2000, managing to withstand the economic crises that have occurred over the years.

It was focused the attention on the units that have not changed the corporate structure. In this way, changes can be considered caused by the structural adjustment, without influence of management changes [2].

The interest in Agricultural sector depends by the grow of economical result even during the economic crisis. In Italy, the added value produced by agriculture has always been increasing, reaching 56.1 billion euros in 2020 [3].

This is the result of socio-economic changes that involved the sector: the Italian farmer cultivates the land to support his family and invests in his funds to create a business [4]. This result in an increase in the average farm size went from 5.5 hectares in 2000 to 8.4 in 2017 [5]. The livestock sector too is affected by a rationalization of resources and a conspicuous increase in intensive farming, leading to a progressive growth in the average size of the animals per farm.

2 Model used and result

The database used is a combination of three sources: the V and V General Census of Agriculture and the Farm Register, the statistical register of Italian farms building up using twelve different sources. According to our hypothesis, it was decided that farms without undergone demographic changes during the period considered would be considered, assuming that a change in owner represents a change in the corporate mission.

The definition of a linkage strategy requires selecting the set of the most discriminant common variables. Linkage methods typically compare different variables of the entities using a set of distance measures. The resulting similarity scores may be combined using different aggregation functions [6].

The V and VI Agricultural Census did not use the same farm identification system. Therefore a statistical matching across the two data sources was necessary.

Multisource approach for trends evaluation. An application at agricultural sector.

The linkage of the statistical units was based on three variables, which identify the farms:

1. Unique Code Farm.
2. Address of the headquarter.
3. Name of the farm.

The first step of the matching model selected for the linkage was to link the Unique Code applying a deterministic model of equality. Then the address and the name were linked by applying a function of the distance of the strings via an indicator normalized between 0 and 1. It measures how information contained in a cell (in this case, the address) is similar to the content of another cell. The value of the index is positively correlated with the degree of similarity in information.

The output resulted was linked with Farm Register using a deterministic model and the Unique Code Farm as matching variable.

Specialization was measured using the Community typology for agricultural holdings (REG EC 1242/2008). It was applied this classification in 2000 and 2010 and calculated the change in the typology for each individual farm.

The result is a panel of 531,536 farms surveyed for the first time in 2000 and listed in the Farm Register in 2017.

2.1 Results

In Italy, the 35% of the farms were already active in 2000. We are talking about 531,136 farms that have been involved in agriculture since 2000 and work a total UAA of 5,525,378.45 hectares.

The territorial distribution of these farms does not differ much from that of 2017, except for a greater presence of farms in the north-east (18% against 14% of the 2017 total) to the detriment of a lower resistance of farms in the South (56% with 60% of the 2017 total).

In particular, Puglia, Sicily, Calabria, Campania, Lazio and Veneto in 2017 held in total almost 60% of Italian farms and 56.4% of panel farms.

Comparing the farms regional share with the panel farms, it is possible to measure the agricultural resilience index according to our definition. The longest-lived farms are located in Northern Italy. The territories with the largest share of 'resilient' farms are Bolzano (46.6%), Friuli-Venezia Giulia (45.1%) and Emilia-Romagna (43, 9%). Of note, for the lowest percentages of long-lived farms, Calabria with a share of 26.7% and Liguria and Lazio with a slightly higher share, equal to 29.7%.

Figure 1 shows the difference between the percentage share of farms by region in 2017 and the percentage share of panel farms. This indicator highlights the differences, not only between north and south, but also between the Tyrrhenian and Adriatic lines where the number of farms survived at the last twenty years is grown.

In addition to a higher agricultural area, the panel farms also started from a different economic strength than the rest of universe: in 2000, 66% of Italian farms had a gross farm product (GFP) of less than 4,800, while of the 500,000 farms considered only 32% had a standard output (SO) of less than 4,000 euros. Although the two methods of calculating the economic value of farms are different between the total number of farms and the panel, it still gives an idea of the phenomenon: in 2000 on average, 1.2% of farms had a GPF of over 120,000 euros, while 2.2% of the panel companies have a SO above 250,000 euros.

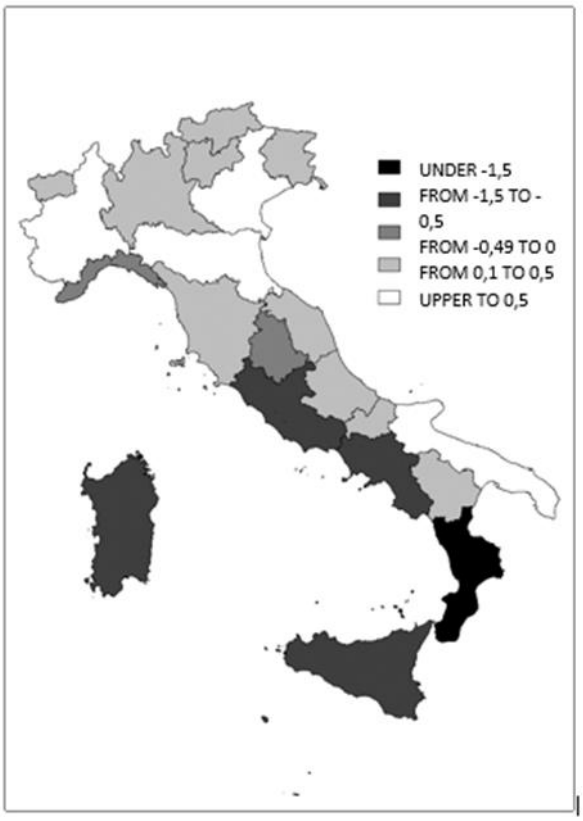


Figure 1 – Differences between resilience index and farms percentage

If these farms had solid premises on which to be able to rely, it is important to observe which production decisions have led them to resist over time. Over the years, in a greater share than the total of Italy, they have lost specialization of mixed farms, oriented to the cultivation of arable land and permanent crops. On the other hand, although there has been a contraction as in the Italian universe, they have remained more anchored to granivores farms. These choices have been successful over time because they have allowed farms, thanks to production changes and adjustments, to

increase their economic value by 23.6%. This is probably the reasons of their resistance to time stability.

3 Conclusion

The value of Italian agri-food system is up to 522 billion and is the one with the highest added value, it emerges that it manages to develop this value from an agricultural area that is half of that of Spain and France [7]. However, the period considered was not without profound structural changes, not least the average size growth in terms of agricultural area. According to some studies [8] this phenomenon requires an adequate study on the reasons connected to this structural change.

The document relay an estimation of the changes in agricultural system using a multisource approach. The results show how this kind of approach could be a solution for solve problems related to the study of structural trends.

References

1. Zhang, L-C., Jentoft, S. (2017). A new GSDEM of multisource data for multiple statistics. UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS Work Session on Statistical Data Editing (den Haag, Netherlands, 24-26 April 2017)
2. Cardillo, C., Fusco, D., Moretti, V., Russo, C. Farms' structural adjustment to the increasing competitive pressure: specialization vs. de-specialization in Italian agriculture. Italian Review of Agricultural Economics **71(1)**, 207-216, (2016).
3. Istat (2021). STIMA PRELIMINARE DEI CONTI ECONOMICI DELL'AGRICOLTURA | ANNO 2020. Statistica report. Rome 22 gennaio 2021.
4. Bianchino, A., Fusco, D., Liguori, M. A. (2020). Struttura e caratteristiche delle aziende agricole, una fotografia tutta italiana. In: Il diritto dell'agricoltura - I/2020. Edizioni scientifiche italiane. Naples
5. Istat (2019). STRUTTURA E CARATTERISTICHE DELLE UNITÀECONOMICHE DEL SETTORE AGRICOLO |ANNO 2017. Statistica report. Rome 2 dicembre 2019.
6. Di Consiglio, L., Fusco D., Tuoto, T., "Exploring solutions for linking Big Data in Official Statistics" Studies in Theoretical and Applied Statistics - Springer Proceedings in Mathematics & Statistics, ISBN 978-3-319-73905-2 2016
7. Crea (2021). Annuario dell'agricoltura italiana 2019. Crea – Centro politiche e bioeconomia. Roma 2017
8. Corsi, A., Di Vita, G., (2017). Cambiamento strutturale dell'agricoltura: il ruolo della demografia e della successione familiare. Agriregionieuropa **anno 13 n°49**, (Giu 2017).

A permutation test on the relationship between Circular Economy and firm size

Test di permutazione sulla relazione tra Economia Circolare e dimensione dell'azienda

Stefano Bonnini and Michela Borghesi

Abstract Circular Economy (CE) has recently become the focus of the debate regarding environmental sustainability. An interesting hypothesis concerns the effect of companies' size on the propensity of SMEs to undertake CE activities. The main difficulty of testing this hypothesis is due to confounding factors such as company age and business sector. We propose a multistrata combined permutation test and we apply it to original data concerning Italian SMEs in the metal sector.

Abstract *L'Economia Circolare (CE) è diventata di recente centrale nel dibattito sulla sostenibilità ambientale. Un'ipotesi interessante riguarda l'effetto delle dimensioni delle imprese sulla propensione delle PMI a intraprendere attività di CE. La principale difficoltà di verificare tale ipotesi riguarda fattori di confondimento come l'età dell'azienda e il settore di attività. Proponiamo un test di permutazione multistrato combinato e lo applichiamo ad un dataset originale sulle PMI italiane nel settore metallurgico.*

Key words: permutation test, nonparametric statistics, confounding effect, stratification, circular economy.

[1] Stefano Bonnini, Department of Economics and Management, University of Ferrara, Via Voltapaletto 11, 44121, Ferrara, Italy; bnnsfn@unife.it

Michela Borghesi, Department of Economics and Management, University of Ferrara, Via Voltapaletto 11, 44121, Ferrara, Italy; michela.borghesi@unife.it

1. Introduction

Circular Economy (CE) provides an important contribution to environmental sustainability mainly (but not only) through the reduction of waste and pollution, the decrease of exploitation of natural resources and the protection of the environment.

Many authors indicates that SMEs are lagging behind (Williamson et al., 2006; Yadav et al., 2018), but they can find challenging to accumulate scientific knowledge and the capabilities required for implementing CE activities and adopting new technologies internally (Ormazabal et al., 2018). Relative to larger firms, SMEs have less access to external finance (Hall et al., 2016), possess lower tangible assets and human capital (Ipinnaiye et al., 2017), and have a smaller market presence (Schot and Steinmueller, 2018). These difficulties often result in SMEs failing to implement or avoiding CE activities altogether (Garcés-Ayerbe et al., 2019; Garrido-Prada et al., 2021).

It seems that firm size (based on number of employees and total turnover) and firms' turnover invested in R&D are significant in explaining within-country variations. Firm size is particularly relevant as medium-sized organizations, both in terms of the number of employees and turnover, are more engaged in CE practices (Hoogendoorn et al., 2015). The decision to undertake CE activities is significantly associated with the number of employees: larger firms are more prone to CE policies (Bassi and Dias, 2019; Bassi and Dias, 2020; Ghisetti and Montresor, 2020).

The main goal of this work is to investigate the relationship between the propensity to undertake CE activities and company size, overcoming the possible confounding effect of firm age. We focus on the metal sector, which is one of the most important in the framework of CE. The metal processing industry has always played a central role in the recovery and recycling of post-consumer metals. Thanks to the recyclability of metals, combined with the versatility that favours their reuse and technologies that increase their durability and lightness, the metallurgical sector has always operated according to an approach of recovery, recycling and reuse, representing a virtuous example of circular economy.

The need of avoiding the possible confounding effects of firm age implies the opportunity of sample stratification and implementation of a multi-strata test. The methodological proposal is based on the application of a nonparametric solution based on a combined permutation test.

Section 2 focuses on the presentation of the statistical problem, followed by the description of the methodological proposal (section 3). The results of the application of the proposed method to a case study, concerning Italian SMEs in the metallurgical sector, are reported in section 4. Section 5 includes concluding remarks.

2. Statistical problem

The main goal of our study is to investigate the effect of firm size on the SMEs' propensity to be involved in CE activities. We deal with a two sample test on

proportions whose H_0 is that the propensity to be circular of medium and small firms is the same. H_1 is that the propensity to be circular of medium firms is greater than that of small firms. We focus on Italian SMEs in the metal sector. The empirical studies that aim to test the effect of a variable on the propensity of enterprises towards CE ignore the possible confounding effects of the economic sector and other factors. Hence, we propose to focus on a specific sector (the metallurgical) and, in order to avoid the possible confounding effect of age, to create strata of firms homogeneous with respect to age, carry out within-stratum tests (comparing medium and small enterprises) and combine the partial tests. Hence, the problem consists in a multistrata test on the proportion for two independent samples.

Let X_j denote the response variable related to the j-th population. X_j takes value 1 if the company undertakes CE activities and 0 otherwise. In our problem $j = 1$ denotes the population of small companies and $j = 2$ the population of medium companies. Under H_0 there is equality in distribution, i.e. all the firms belong to a unique population. Under H_1 there is a stochastic dominance of X_2 (X_2 dominates X_1). In other words, H_1 is equivalent to: $P(X_1 = 1) < P(X_2 = 1)$ or equivalently $\theta_1 < \theta_2$ with $\theta_j = P(X_j = 1)$. Equivalently, within the s-th stratum

$$\begin{cases} H_{0s}: X_{s1} \overset{d}{=} X_{s2} \\ H_{1s}: X_{s1} \overset{d}{<} X_{s2} \end{cases}$$

where X_{sj} denotes the response variable for the s-th stratum (age group) and the j-th population (size group).

The general problem can be defined according to the union/intersection approach. Hence, the multi-stratum test is

$$\begin{cases} H_0: \bigcap_s H_{0s} \\ H_1: \bigcup_s H_{1s} \end{cases}$$

because the null hypothesis is true if all the partial null hypotheses are true and the alternative hypothesis is true if at least one partial alternative hypothesis is true.

3. Methodological solution

The proposed solution is based on the application of a combined permutation test (Pesarin and Salmaso, 2010). This family of nonparametric methods is suitable when the problem can be broken down into k sub-problems or partial tests. Hence, we have k partial null hypotheses $H_{01}, \dots, H_{0k}$ and k partial alternative hypotheses $H_{11}, \dots, H_{1k}$ and the overall problem can be defined as

$$\begin{cases} H_0: \bigcap_{i=1}^{k} H_{0i} \\ H_1: \bigcup_{i=1}^{k} H_{1i} \end{cases}$$

Basically, it is a multiple test where each sub-problem consists in testing the null hypothesis H_{0i} versus the alternative hypothesis H_{1i}. Permutation methods can be used, provided that mean and variance of the populations are assumed to be finite and exchangeability under the null hypothesis holds (Pesarin, 2001). Permutation methods are preferable to parametric solutions when the underlying distribution is unknown or cannot be assumed according to asymptotic theories (hence especially for small samples). Moreover, in the presence of multiple tests such as the one considered, the dependence between the test statistics of the partial problems does not need be explicitly modeled, as in the likelihood approach or other parametric methods. In particular, with the combined permutation tests, the dependence structure is implicitly taken into account by permuting the rows of the dataset and the application of the combining function ψ. The sufficient statistic of permutation tests is represented by the observed dataset.

Without loss of generality, we can assume that the null (partial and overall) hypotheses are rejected for large values of the test statistics. Let $L_i(t) = P(T_i \geq t|\boldsymbol{X})$ denote the significance level function of the i-th partial test, T_i the test statistic of the i-th partial test and t_i a given value taken by T_i. The (univariate) combined test statistic is $T_\psi = \psi(l_1, \cdots, l_k)$ where $l_i = L_i(t_i)$. ψ must be non-increasing function of the arguments and satisfy mild conditions such as: it tends to its supremum (possibly not finite) when one argument tends to zero and, $\forall \alpha \in (0,1)$, the critical value of T_ψ is assumed to be finite and strictly less than the supremum.

For the problem under study, the within-stratum tests represent the sub-problems and a suitable partial test statistic is the difference of sample proportions. Let $\hat{\theta}_{sj} = p_{sj} = f_{sj}/n_{sj}$ be the sample proportion of circular companies in the j-th sample of the s-th stratum, with f_{sj} and n_{sj} absolute frequency of circular companies and sample size respectively in the j-th sample of the s-th stratum. Thus, the partial test statistic is $T_s = \hat{\theta}_{s1} - \hat{\theta}_{s2}$. A possible combination that provides powerful tests when the number of true partial alternative hypotheses is low, is given by

$$T_\psi = max(1 - l_1, \cdots, 1 - l_k)$$

Such combined test is exact, unbiased, consistent and distribution free.

4. Case study

The research question of the case study is: "Is the propensity towards CE of small firms less than that of medium firms, taking into account the possible confounding effect of firm age?". In order to answer this question, we applied the combined permutation test with stratification described above. The application taken into account concerns the Italian SMEs in the metal sector. The dataset is original and

related to a sample survey about CE on Italian SMEs carried out by telephone interview in January 2020. The sample consisted in 475 Italian firms operating in the metal sector. The variable of interests in the problem are three:

- response variable (dummy): it indicates the firm propensity towards CE (1: yes, 0: no). It corresponds to the question "Did you make investments in R&D aimed at reducing the environmental impact of production?"
- factor (dummy): it denotes firm size (0: less than 16 employees, 1: 16 or more employees) and represents the "treatment". The goal is to test the significance of the factor's effect on the response.
- confounder (categorical): it denotes firm age (1: less than 17 years old, 2: from 17 to 36 years old, 3: from 37 to 56 years old, 4: more than 56 years old).

In Table 1 the sample proportions of circular firms within the groups of small and medium SMEs for each age group are reported. It is evident that, for the first two age groups, the propensity to be circular is greater in small companies while for older companies there are similar proportions.

Table 1: Sample proportions of firms that made investments in R&D aimed at reducing the environmental impact of production.

	Firm size	
Firm Age	< 16 employees	≥ 16 employees
< 17 years	0.019	0.004
$17 - 36$ years	0.021	0.008
$37 - 56$ years	0.015	0.013
> 56 years	0.006	0.006

By carrying out the combined permutation test, we obtained an overall p-value of 0.645, which indicates no significance at $\alpha = 0.05$. Hence, there is not empirical evidence to reject the hypothesis of null effect of firm size in favor of the alternative hypothesis that medium enterprises have a greater propensity towards CE.

5. Concluding remarks

In the empirical literature on Circular Economy, there is a lack of contributions on the effects of firm size on the propensity towards CE. The few existing works adopt approaches that do not take into account some typical confounding effects, such as those of the economic sector and of the company's age. We propose the application of a multistrata procedure based on a combined permutation test.

This methodology permits to test complex hypotheses, it is powerful, flexible because distribution-free, and satisfy important properties such as unbiasedness and

consistency. This approach overcomes the limitations of the parametric methods usually applied to such problems, based on restrictive and unrealistic assumptions.

The application of the test to original sample data concerning Italian SMEs in the metal sector does not bring to empirical evidence in favor of the hypothesis that firm size affects the propensity to Circular Economy.

Acknowledgements

The work was supported by the Italian Ministry of Education, University & Research that funded the departmental development program (DEM – University of Ferrara) for the period 2018–2022, to promote excellence in education and research ("Dipartimenti di Eccellenza").

References

1. Bassi, F., Dias, J.G.: The use of circular economy practices in SMEs across the EU. Conservation & Recycling **146**, 523–533 (2019).
2. Bassi, F., Dias, J.G.: Sustainable development of small- and medium-sized enterprises in the European Union: A taxonomy of circular economy practices. Business Strategy and the Environment **29**, 2528–2541 (2020).
3. Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric Hypothesis Testing, rank and permutation methods with applications in R. Wiley series in probability and statistics, Chichester (2014).
4. Garcés-Ayerbe, C., Rivera-Torres, P., Suàrez-Perales, I., Leyva-de la Hiz D.I.: Is it possible to change from a linear to a circular economy? An overview of opportunities and barriers for European small and medium-sized Enterprise companies. Int. J. Environ. Res. Public Health **16**(5), 851 (2019).
5. Garrido-Prada, P., Lenihan, H., Doran, J., Rammer, C., Perez-Alaniz, M.: Driving the circular economy through public environmental and energy R&D: Evidence from SMEs in the European Union. Ecological Economics **182**, 106884 (2021).
6. Ghisetti, C., Montresor, S.: On the adoption of circular economy practices by small and medium-size enterprises (SMEs): does "financing-as-usual" still matter?. Journal of Evolutionary Economics **30**, 559–586 (2020).
7. Hall, B.H., Moncada-Paternò-Castello, P., Montresor, S., Vezzani, A.: Financing constraints, R&D investments and innovative performances: new empirical evidence at the firm level for Europe. Econ. Innov. New Technol **25** (3), 183–196 (2016).
8. Hoogendoorn, B., Guerra, D., van der Zwan, P.: What drives environmental practices of SMEs?. Small Bus Econ **44**, 759–781 (2015).
9. Ipinnaiye, O., Dineen, D., Lenihan, H.: Drivers of SME performance: a holistic and multivariate approach. Small Bus. Econ **48**(4), 883–911 (2017).
10. Ormazabal, M., Prieto Sandoval, V., Puga-Leal, R., Jaca, C.: Circular economy in spanish SMEs: challenges and opportunities. J. Clean Prod **185**, 157–167 (2018).
11. Pesarin, F., Salmaso, L.: Permutation tests for complex data: theory, applications and software. Wiley, Chichester (2010).
12. Pesarin, F.: Multivariate permutation tests with applications in biostatistics. Wiley, Chichester (2001).
13. Schot, J., Steinmueller, W.E.: Three frames for innovation policy: R&D, systems of innovation and transformative change. Res. Polic. **47** (9), 1554–1567 (2018).
14. Williamson, D., Lynch-wood, G., Ramsay, J.: of Environmental SMEs Behaviour and the in Manufacturing Implications for CSR. Journal of Business Ethics **67**(3), 317–330 (2006).
15. Yadav, N., Gupta, K., Rani, L., Rawat, D.: Drivers of Sustainability Practices and SMEs: A Systematic Literature Review. European Journal of Sustainable Development **7**(4), 531-544 (2018).

Circular economy and business models: a literature review

Economia circolare e modelli di business: la revisione della letteratura

Stefania Mele, Filomena Izzo, Viktoriia Tomnyuk[1]

Abstract Attention to the circular economy is growing in government, business, society, and academia. However, the transition to a circular economy requires the adaption of business models or the creation of new ones. Unfortunately, the current literature has lacked a consolidated understanding of the current knowledge on circular business models. The study wants to investigate how the circular economy and business models are related in the current literature through a bibliometric analysis—investigating the dominant themes of research in CE and business models and the avenues for future research.

Abstract *L'attenzione all'economia circolare sta crescendo nelle istituzioni pubbliche, nelle imprese, nella società e nel mondo accademico. Tuttavia, la transizione verso un'economia circolare richiede l'adeguamento dei modelli di business o la creazione di nuovi modelli. Nella letteratura attuale manca una comprensione consolidata dello stato attuale delle conoscenze sui modelli di business circolari. Lo studio vuole indagare come l'economia circolare e i modelli di business sono correlati nella letteratura attuale attraverso un'analisi bibliometrica. Indagando, in particolare, i temi dominanti della ricerca in economia circolare e modelli di business e quali sono le strade per la ricerca futura.*

Keywords: Circular economy, business, literature review, bibliometrix.

[1] Filomena Izzo, University of Campania "Luigi Vanvitelli", Economics Department, Capua (CE), Italy; filomena.izzo@unicampania.it:

Stefania Mele University of Campania "Luigi Vanvitelli", Economics Department, Capua (CE), Italy; stefania.mele@unicampania.it

Viktoriia Tomnyuk, Department of Cultures, Politics and Society, Università degli Studi di Torino, Turin, Italy; viktoriia.tomnyuk@unito.it.

1 Introduction

In the last years, it has been widely recognized that shifting from the linear economy model to a circular one brings environmental, social, and financial advantages [8,6]. In a circular economy (CE), the economic and environmental value of materials is preserved for as long as possible by keeping them in the economic system, either by extending the life of the products created from them or by looping them back into the system to be reused [5].

Given its substantial impact on the environment, the circular economy has become a fundamental topic in public debates. The EU declares the necessity for academic research on innovative and more sustainable economic models and strategies [11,6]. The transition to the circular economy often requires holistic adaptations in firms' business models or even the creation of new ones [2,9,6].

The standard definition of Business Models can be outlined as the configuration of customer sensing, customer engagement, value delivery, and monetization components that capture causal links between value creation and value capture at the business level [12,1,4].

Despite the growing attention from the literature, we currently lack a consolidated understanding of the current state of knowledge on circular business models. In particular, because many studies were published in a short period, their structures and discourses are not well established and linked to each other [6]. Furthermore, environmental and engineering sciences have contributed the most to the CE literature in fields such as industrial ecology compared to management studies [7,10,3]. There is little engagement among management and organization studies scholars with the CE [3].

The work is structured as follows: First, the articles are analyzed according to network and cluster analysis principles, using the Bibliometrix R package to explore the most-researched terms and their relationships and identify less-explored terms and research gaps. Then will be furthermore conducted a qualitative review of selected publications to illustrate quantitative results and examine more profound research topics.

1.1 Research Questions

The study wants to investigate how the circular economy and business models are related in the current literature. With this objective, the paper addresses the following research questions.

1. How has the research landscape of circular economy and business models domain evolved?
2. Which authors and articles are the most influential in circular economy and business models?
3. Which institution and country are the most influential?
4. What are the dominant themes of research in circular economy and business models?
5. What are the collaboration patterns between the authors contributing to the circular economy and business models?
6. How have specific keywords and themes evolved from 2011 to 2021?
7. What are the avenues for future research?

2 Research methodology

The bibliometric analysis search was conducted in September 2021 and utilized the Web of Science (WoS) database. Compared with other databases (e.g., Google Scholar and Scopus) with more publications, WoS is considered a leading quality-oriented database in the academic world. We used a keyword search with "circular economy" AND "business model" OR "circular business model." To identify the most influential work in the business model and circular economy field, we focused on journal articles since formally subjected to a double-blind peer-review process to guarantee a high standard of papers included in the final sample. A total of 264 items in WoS matched the criteria and were included for further examination.

Furthermore, we manually reviewed the titles, abstracts, and keywords of the selected articles, double-checking for the selection criteria and deleting those sources that were not in the field of business and management and/or did not deal with the circular economy and business models.

References

1. Baden-Fuller, C., & Mangematin, V. Business models: A challenging agenda. Strategic Organization. 11(4), 418-427 (2013)
2. Bocken, N. M., De Pauw, I., Bakker, C., & Van Der Grinten, B. Product design and business model strategies for a circular economy. Journal of industrial and production engineering. 33(5), 308-320 (2016)
3. De Angelis, R. Circular economy: Laying the foundations for conceptual and theoretical development in management studies. Management Decision (2020).
4. De Giacomo, M. R., & Bleischwitz, R. Business models for environmental sustainability: Contemporary shortcomings and some perspectives. Business Strategy and the Environment, 29(8), 3352-3369 (2020)
5. Den Hollander, M. C., Bakker, C. A., & Hultink, E. J. Product design in a circular economy: Development of a typology of key concepts and terms. Journal of Industrial Ecology, 21(3), 517-525 (2017)
6. Ferasso, M., Beliaeva, T., Kraus, S., Clauss, T., & Ribeiro-Soriano, D. Circular economy business models: The state of research and avenues ahead. Business Strategy and the Environment, 29(8), 3006-3024 (2020)
7. Lahti, T., Wincent, J., & Parida, V. A definition and theoretical review of the circular economy, value creation, and sustainable business models: where are we now and where should research move in the future?. Sustainability, 10(8), 2799 (2018)
8. Lewandowski, M. Designing the business models for circular economy—Towards the conceptual framework. Sustainability, 8(1), 43 (2016)
9. Manninen, K., Koskela, S., Antikainen, R., Bocken, N., Dahlbo, H., & Aminoff, A. Do circular economy business models capture intended environmental value propositions?. Journal of Cleaner Production, 171, 413-422 (2018)
10. Sehnem, S., Vazquez-Brust, D., Pereira, S. C. F., & Campos, L. M. Circular economy: benefits, impacts, and overlapping. Supply Chain Management: An International Journal (2019)
11. Urbinati, A., Chiaroni, D., & Chiesa, V. Towards a new taxonomy of circular economy business models. Journal of Cleaner Production, 168, 487-498 (2017)
12. Zott, C., Amit, R., & Massa, L. The business model: recent developments and future research. Journal of Management, 37(4), 1019-1042 (2011)

Session of free contributes SCL11 –*Modelling Extreme Values, High dimensional, time series data*
Chair: Violetta Simonacci

Modeling extreme values using the r-largest four parameter distribution

Yire Shin, Piyapatr Busababodhin, Jeong-Soo Park*

Abstract The generalized extreme value distribution (GEVD) has been widely used to model the extreme events in many areas. It is however limited to using only block maxima, which motivated to model the GEVD dealing with r-largest order statistics (rGEVD). The rGEVD which uses more than one extreme per block can significantly improves the performance of the GEVD. The four parameter kappa distribution (K4D) is a generalization of some three-parameter distributions including the GEVD. It can be useful in fitting data when three parameters in the GEVD are not sufficient to capture the variability of the extreme observations. The K4D still uses only block maxima. In this study, we thus extend the K4D to deal with r-largest order statistics as analogy as the GEVD is extended to the rGEVD. The new distribution is called the r-largest four parameter kappa distribution (rK4D).

Key words: r-largest order statistics, Hydrology, Annual maximum sea level.

1 Introduction

The generalized extreme value distribution (GEVD) has been widely used to analyse univariate extreme values (Coles 2001). The GEVD encompasses all three possible asymptotic extreme value distributions predicted by large sample theory. The cumulative distribution function (cdf) of the GEVD is as follows (Hosking and Wallis 1997):

Yire Shin, Department of Mathmatics & Statistics, Chonnam National University, South Korea; email:shinyire@hanmail.net

Piyapatr Busababodhin, Department of Mathmatics & Statistics, Chonnam National University, South Korea; email:jspark@chonnam.ac.kr

Jeong-Soo Park, Department of Mathmatics, Mahasarakham, Thailand

$$F_3(x) = exp\left\{-(1 - k\frac{x-\mu}{\sigma})^{1/k}\right\}$$

When $1 - k(x-\mu)/\sigma > 0$ and $\sigma > 0$, where μ, σ, k are the location, scale, and shape parameters, respectively. The particular case for k = 0 in (1) is the Gumbel distribution. Note that the sign of k is changed from the book of Coles (2001).

One difficulty of applying the GEVD is using the limited amount of data for model estimation. Since extreme values are scarce, making effective use of the available information is important in extremes. This issue has motivated the search for a model to use more data other than just block maxima. The inclusion of more data up to r-th order statistics in each block other than just maxima will improve precision of model estimation, but the interpretation of parameters is unaltered from the univariate GEVD for block maxima. The above univariate result was extended to the r-largest order statistics model, which gives the joint density function of the limit distribution (Coles 2001);

$$f_3\left(\underline{x}^{(r)}\right) = \exp\left\{-\omega(x^{(r)})^{1/k}\right\} \times \prod_{s=1}^{r} \sigma^{-1}\omega(x^{(s)})^{\frac{1}{k}-1}$$

$$where\ x^{(1)} \geq x^{(2)} \geq \cdots \geq x^{(r)},\ and\ w\left(x^{(s)}\right) = 1 - k\frac{x^{(s)}-\mu}{\sigma} > 0\ for\ s = 1,2,\cdots,r$$

The rGEVD was encouraged to use by Zhang (2004), and has been employed in some real applications (Soares and Scotto 2004; An and Pandey 2007; Wang and Zhang 2008; Feng and Jiang 2015; Naseef and Kumar 2017). The number r comprises a bias-variance trade-off: small values of r generate few data leading to high variance; large values of r are likely to violate the asymptotic support for the model, leading to bias (Coles 2001). Bader et al.(2017) developed automated methods of selecting r from the rGEVD.

The inclusion of more data up to r-th order statistics in each block other than just maxima will improve precision of model estimation, but the interpretation of parameters is unaltered from the univariate GEVD for block maxima. For small to moderate sample sizes, the GEVD sometimes yields inadequate results. It may be because the GEVD is derived by a large sample theory for the extremes of independent sequences.

As a generalization of some common three-parameter distributions including the GEVD, the four parameter kappa distribution (K4D) was introduced by Hosking (1994). It can be useful in fitting data when three parameter distributions including the GEVD are not sufficient to capture the variability of observations. Some researchers studied on the K4D (Dupuis 1997; Dupuis and Winchester 2001; Singh and Deng 2003; Park and Kim 2007; Murshed et al. 2014).

The probability density function (pdf) of K4D is,

$$f_4(x) = \sigma^{-1}\omega(x)^{(1/k)-1}F_4(x)^{1-h}$$

$where\ w(x) = 1 - k\frac{x-\mu}{\sigma}$, $\mathrm{F_4(x)} = \left\{1 - \mathrm{h}\omega(\mathrm{x})^{1/\mathrm{k}}\right\}^{1/\mathrm{h}}$ is the cdf of the K4D.

The K4D includes many distributions as special cases, as shown in Figure1 the generalized Pareto distribution for h=1, the GEVD for h=0, the generalized logistic distribution for h=-1, the generalized Gumbel distribution for k=0, the Gumbel distribution for h=0, k=0. The K4D is flexible and widely applicable to the data including not only extreme values but also skewed data. It has been used in many fields, particularly in hydrology and atmospheric sciences, for fitting extreme values or skewed data (e.g., Parida 1999; Park and Jung 2002; Seo et al. 2015; Kjeldsen et al. 2017; Brunner et al. 2019; Jung and Schindler 2019). Hosking and Wallis (1997) employed the K4D in regional frequency analysis as a parent distribution from which the samples are drawn. Blum et al.(2017) found that the K4D provides a very good representation of daily streamflow across most physiographic regions in the conterminous United States

In analyzing extreme values, the K4D has the same limitation of using only the block maxima as the GEVD has like as the GEVD was extended to rGEVD, an extension of the K4D to r-largest order statistic model may be very useful to address this limitation. The inclusion of more observations up to r-th order statistics other than just maxima will improve precision of model estimation. The extension in the K4D is not published yet.

In this study, we thus developed an r-largest order statistics model as an extension of the K4D as well as of the rGEVD. It is referred to the rK4D. Figure2 illustrates our motivic schema. The remainder of this paper is organized as follows. Section 2 includes the definition of the rK4D. Section 3 some practical concerns of the rK4D by applying it to Bangkok rainfall data. Section 4 concludes with discussion.

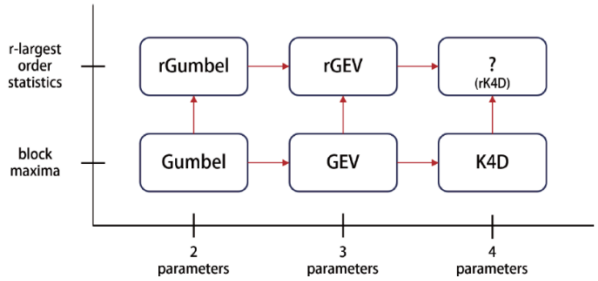


Figure 1. A motivic schema on generalizations from 2 parameters to 4 parameters, and extensions from the block maxima models to the r-largest order statistic models, which leads to the r-largest four parameter kappa distribution (rK4D).

2 r-largest four parameter kappa distribution

The r-largest four parameter kappa distribution (rK4D) is not the result from any theoretical derivation but just an analogous extension from the K4D and the rGEVD. To define the joint probability density function (pdf) of the rK4D, we considered and

followed the generalization processes from the GEVD to the K4D and to the rGEVD. We define the joint pdf of the Rk4d; under k ≠0, h ≠0,

$$f_4\left(\underline{x}^{(r)}\right) = \rho^{-r} C_r \times g(\underline{x}^r) \times F_4(x^{(r)})^{1-rh}$$

$$C_r = \begin{cases} \prod_{i=1}^{r-1} [1-(r-i)h] & if\ r \geq 2 \\ 1 & if\ r = 1 \end{cases}$$

$$g\left(\underline{x}^r\right) = \prod_{s=1}^{r} \omega(x^{(s)})^{\frac{1}{k}-1}$$

The supports of this pdf are $x^{(1)} \geq x^{(2)} \geq \cdots \geq x^{(r)}$, $\sigma > 0$, $w\left(x^{(s)}\right) > 0\ for\ s = 1,2,\cdots, r,\ C_r > 0,\ and\ 1 - h \times w\left(x^{(r)}\right)^{1/k} > 0$, When $r = 1$, this pdf is same as the pdf of the K4D in (3), when $h \to 0$, this pdf goes to the pdf of the rGEVD.

3 Real application : Bangkok data

The top 10 annual rainfall events were taken from the daily records of a rain gange station in Bangkok City, from 1960-2018. The rK4D model is fitted to the values for r=1,2,…,10. The MLE of parameters and the 20-year return levels with standard errors in the parenthesis for several values of r are given in Table 1. For comparison, similar results from the fitted rGEVD are also presented. The upper table is for the rGEVD and the lower one is for the rK4D. In Table 1, the standard errors of parameter estimates decrease with increasing values of r for the rGEVD. That is not obvious in the rK4D but generally shows a decreasing trend. These non-monotonic decreasing cases may be because of the trouble in numerical optimization with 4 parameters in the rK4D or the intrinsic property of the rK4D. The SEs of $\hat{\mu}$, $\hat{\sigma}$, and $\hat{k}$ in the rK4D are generally bigger than those in the rGEVD. The SEs of h estimates in the rK4D are much larger compared to those of the other parameter estimates.

Table 1: The estimates of parameters and 20-year return level (r20) with standard errors (se) of the estimates in parenthesis which are obtained from the r-largest order statistic models fitted to Bangkok rainfall data with different values of r. Upper table is for the rGEVD and lower one is for the rK4D. `nllh' stands for the negative log-likelihood function value.

r	nllh	$\hat{\mu}$ (se)	$\hat{\sigma}$ (se)	$\hat{k}$ (se)	rGEV r_{20}(se)
1	293.8	90.6 (4.2)	28.3 (3.2)	0.098(0.114)	188.2 (18.3)
2	517.5	89.4 (3.3)	26.8 (2.5)	0.153(0.091)	190.4 (20.0)
3	708.1	90.6 (3.1)	27.0 (2.3)	0.122(0.069)	187.3 (17.2)
4	871.2	91.0 (2.9)	26.6 (2.2)	0.126(0.061)	186.8 (16.7)
5	1028.9	91.6 (2.9)	26.5 (2.1)	0.095(0.052)	182.6 (14.6)
6	1170.6	91.9 (2.8)	26.4 (2.0)	0.088(0.048)	181.5 (14.0)

Modeling extreme values using the r-largest four parameter distribution

7	1297.5	91.7 (2.8)	26.3 (2.0)	0.093(0.045)	181.6 (13.9)
8	1417.9	91.7 (2.8)	26.3 (2.0)	0.096(0.043)	182.0 (13.9)

r	nllh	$\hat{\mu}$ (se)	$\hat{\sigma}$ (se)	$\hat{k}$ (se)	$\hat{h}$	$rK4Dr_{20}$(se)
1	293	78.7 (11.4)	43.5 (14.8)	0.12 (0.170)	0.527 (0.30)	188.0 (13.7)
2	517.4	89.4 (3.3)	28.0 (3.3)	-0.11 (0.124)	0.073 (0.13)	187.9 (18.7)
3	708	90.4 (3.1)	26.5 (3.1)	-0.15 (0.090)	-0.043 (0.10)	189.3 (18.9)
4	870.5	90.3 (3.1)	26.0 (3.1)	-0.18 (0.076)	-0088 (0.09)	191.9 (19.7)
5	1028.5	90.9 (3.0)	26.1 (3.0)	-0.14 (0.063)	-0.068 (0.06)	186.3 (16.7)
6	1169.6	91.0 (3.0)	26.2 (3.0)	-0.13 (0.057)	-0.601 (0.05)	185.6 (16.1)
7	1296.2	90.9 (2.9)	26.3 (2.9)	-0.13 (0.054)	-0.057 (0.04)	186.7 (16.3)
8	1416.7	91.0 (2.9)	26.4 (2.9)	-0.13 (0.051)	-0.046 (0.04)	186.8 (16.0)

The 20-year return levels and its standard errors (SE) decrease with r in rGEVD, whereas those values for rK4D do not show a monotonic decrease. This phenomenon for the return levels of the rK4D is probably explained by that the return level and its SE are obtained for the annual maximum while the rK4D is fitted to the r-largest order statistics. Because the parameter estimates of the rK4D are obtained to take account into all data up to the r-largest observations, it may not work good for the annual maximum only.

This phenomenon may be more serious for the rK4D than the rGEVD because the standard errors of the return levels of the rK4D are greater than those of the rGEVD. This is a re-confirmation of the general rule that the model with more parameters usually results in bigger variance (and less bias) than the model with fewer parameters (James et al. 2013).

4 Conclusion and discussion

In this study, we introduced the r-largest four parameter kappa distribution (rK4D). Application to Bangkok rainfall data is presented with comparison to the r-largest GEVD. This study illustrates that the rK4D gives better fitting or less biases but larger variances of the parameter estimates than the rGEVD. The pdf definition of the rk4d may not be unique, because it is not a result from any theoretical derivation but just an analogous extension from the K4D and the rGEVD. A point process approach for extremes (Smith 1989; Coles 2001) may provide a theoretical insight. The rK4D, as an extension of the rGEVD, can serve to model the r-largest observations flexibly with less bias than the rGEVD, specially when three parameters in the rGEVD are not enough to capture the variability of observations well. Even though there are defects such as larger estimation variance in the rK4D compared to

the rGEVD, the introduction of the rK4D will enrich and improve our modelling methodology for extreme events.

Acknowledgments: This work was supported by the BK21 FOUR funded by the Ministry of Education, Korea (No. 5120200913674) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (No.2020R1I1A3069260, No. 2021R1A6A3A13044162).

References

1. An, Y., Pandey, MD. : The r largest order statistics model for extreme wind speed estimation. Journal of Wind Engineering and Industrial Aerodynamics **95** 165-182. (2007)
2. Bader, B., Yan, J., Zhang, XB. : Automated selection of r for the r largest order statistics approach with adjustment for sequential testing. Statistics and Computing **27** 1435-1451. (2017)
3. Coles, S. : An introduction to statistical modeling of extreme values. Springer, London, pp 224 (2001)
4. Hosking, JRM. : The four-parameter kappa distribution. IBM Journal of Research and Development **38** 251-258. (1994)
5. Hosking, JRM., Wallis, JR. : Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge. pp 244. (1997)
6. Park, JS., Kim, TY. : Fisher information matrix for a four-parameter kappa distribution. Statistics & Probability Letters **77(13)** 1459–1466. (2007)
7. Singh, VP., Deng, ZQ. Entropy-based parameter estimation for kappa distribution. Journal of Hydrologic Engineering **8** 81-92. (2003)
8. Soares, CG., Scotto, MG. : Application of the r largest-order statistics for long-term predictions of significant wave height. Coastal Engineering **51** 387-394. (2004)
9. Zhang, XB., Zwiers, FW., Li, GL. : Monte Carlo experiments on the detection of trends in extreme values. Journal of Climate **17** 1945-1952. (2004)

Classification of ECG signals based on functional data analysis and machine learning techniques

Classificazione dei segnali ECG basata sull'analisi dei dati funzionali e tecniche di apprendimento automatico

Mohammed Sabri, Fabrizio Maturo, Rosanna Verde, Jamal Riffi, Ali Yahyaouy and Hamid Tairi

Abstract High-dimensional data classification is always a challenging task due the so-called curse of dimensionality issue. This study proposes a two-steps supervised classification technique for high-dimensional time series treated as functional data. The first phase is based on the idea of extracting additional knowledge from the data using unsupervised classification by means of a new distance that considers the original curves and their derivatives. The second step involves functional supervised classification of the new patterns discovered. Particularly, a Random Forest classifier is built using the new labels obtained in the first step. The experiments on ECG data and comparison with the classical approaches show the effectiveness and exciting improvement in terms of accuracy.

Abstract La classificazione dei dati ad alta dimensionalità è un problema complesso a causa della cosiddetta maledizione della dimensionalità. Questo studio propone una tecnica di classificazione supervisionata in due fasi per serie temporali ad alta dimensionalità trattate come dati funzionali. La prima fase si basa sull'idea di estrarre informazioni dai dati utilizzando una classificazione non supervisionata basata su una nuova distanza. La seconda fase prevede la classificazione supervisionata funzionale considerando i nuovi la-

Mohammed Sabri
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy.
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, e-mail: sabri.mohammed@unicampania.it

Fabrizio Maturo, Rosanna Verde
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: fabrizio.maturo@unicampania.it, rosanna.verde@unicampania.it

Jamal Riffi, Ali Yahyaouy, Hamid Tairi
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco e-mail: riffi.jamal@gmail.com, ali.yahyaouy@usmba.ac.ma, htairi@yahoo.fr

bels scoperti. In particolare, viene utilizzato il random forest utilizzando le nuove etichette. Gli esperimenti sui dati ECG e il confronto con i classici approcci di classificazione funzionale mostrano l'efficacia del metodo e un ottimo miglioramento dell'accuratezza del classificatore.

Key words: Functional Data Analysis (FDA), supervised classification, Functional K-means, Functional Random Forest

1 Introduction

The main objective of the electrocardiogram (ECG) analysis is to detect the life-threatening arrhythmias accurately for appropriate treatment in order to save life. During the last decades, several methods were reported for automatic ECG beat classifications. ECG data is a classic example of high-dimensional data, and therefore their classification requires adequate statistical techniques. Classification of high-dimensional data is a fast-growing research area, driven by a need for methods to deal with the increasing availability of data coming from sensors and biomedical devices. However, due to the course of dimensionality problem, it is always challenging to build an accurate model that is reasonably flexible and yet feasible to fit.

A possible approach to deal with high-dimensional data is functional data analysis (FDA), i.e. considering each time series as a single entity given by a functional object [1]. In this context, recently, several metrics and semimetrics have been proposed to compute the similarity among curves, and many approaches have been suggested to deal with functional data classification problems.

Our starting idea to improve the classical functional classifiers' performance is to consider ECG data as functional data in the time domain [3, 4] and propose an original two-phase classification approach. In the first step, a functional clustering algorithm is used to discover new patterns in the original classes, e.g. the functional K-means algorithm. Then, supervised classification based on Random Forest (RF) is applied exploiting the additional information on the new labels coming from the first step. The basic idea is to get additional knowledge in the training data to improve the power of the final classifier. Also, we define a novel distance used to measure the similarity among functional samples by considering also the information of their derivatives.

2 The two-phases functional classification procedure

FDA analyses samples where each observation arises from a function varying over a continuum. For ECG data, the continuum is function. Thus, we consider a set X of ECG signals where y_i $(i = 1, \ldots, N)$ indicate the class label of the $i-th$ signal. The basic idea is to represent each ECG signal as a functional data that can be expressed as a linear combination of basis functions, e.g. b-splines. B-spline. Assuming fixed basis, each i-th signal can be approximated as follows:

$$x_i(t) = \sum_{j=1}^{p} c_{i,j}\psi_j(t) \tag{1}$$

where ψ_j are p known basis functions and $c_{i,j}$ are the corresponding coefficients to be estimated. As the ECG data is usually a non-periodic data, B-spline basis system are used. The B-splines basis coefficients are estimated by the ordinary least squares method, minimizing the sum of squared residuals [1, 2].
The first advantage of using FDA is that we consider the whole shape of the curves to compute the similarity among functional data and classify them. The second advantage is that we can deal with the curse of dimensionality issue by using a low number of coefficients via dimensionality reduction techniques able to create independent features.
The main idea of our proposal is using clustering to discover new patterns in the dataset, and thus exploiting a combination of supervised and unsupervised methods to improve the performance of a functional classifier.
In addition, this research proposes a new distance to measure the similarity between two functional data. The latter semi-metric also involves the use of derivatives to consider the similarity between curves to take into account additional behaviours of the functions.
Given two functional data $x_i(t)$ and $x_j(t)$ from a data set X, a new similarity metric between $x_i(t)$ and $x_j(t)$ is defined as

$$d^2(x_i(t), x_j(t)) = d_{ij}^{(0)} + d_{ij}^{(1)} + d_{ij}^{(2)} \tag{2}$$

with
$d_{ij}^{(0)} = \frac{1}{\int_T \sigma_{x(t)}(dt)} \int_T (x_i(t) - x_j(t))^2 dt$
$d_{ij}^{(1)} = \frac{1}{\int_T \sigma_{Dx(t)}(dt)} \int_T (Dx_i(t) - Dx_j(t))^2 dt$
$d_{ij}^{(2)} = \frac{1}{\int_T \sigma_{D^2x(t)}(dt)} \int_T (D^2x_i(t) - D^2x_j(t))^2 dt$
and where $Dx_i(t)$ is the first-order functional derivative of the i-th curve, and $D^2x_i(t)$ is the second-order functional derivative of the i-th curve.

Silhouette analysis [8] is used to determine the most suitable number of subgroups of each original label by using the new distance defined in Equation (2); the functional k-means clustering algorithm, based on the distance defined in Equation (2), is implemented to discover the functional elements of each subgroup in the training set by employing the number of clusters we get from the silhouette method. The entire process of the functional k-means clustering algorithm based on the distance d in Equation (2) is described as in Algorithm 1.

Algorithm 1 The functional k-means clustering algorithm based on the new metric

input:
- X_g group of curves in the group g ($g = 1, ..., G$)
- K_g number of clusters (subgroups) found in the group g

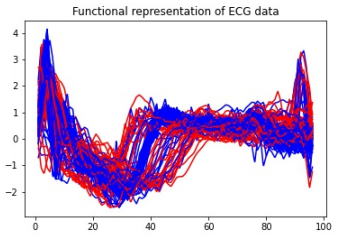
Output: - $labels_g$: the new labels of the group X_g

for g $\leftarrow$ 1 to G **do**
  Randomly choose K_g samples as the initial centroids $\mu_1(t), \ldots, \mu_{K_g}(t)$
  while stopping criterion has not been met **do**
    for i $\leftarrow$ 1 to K_g **do**
      for $x(t) \in X_g$ **do**
        $j \leftarrow \arg\min_i \ d(\mu_i(t), x(t))$
        $G_j \leftarrow G_j \bigcup x(t)$
      end for
      for m $\leftarrow$ 1 to K_g **do** $\mu_m(t) \leftarrow \frac{1}{|G_m|} \sum_{x(t) \in G_m} x(t)$
      end for
    end for
  end while
  $labels_g \leftarrow G$
end for

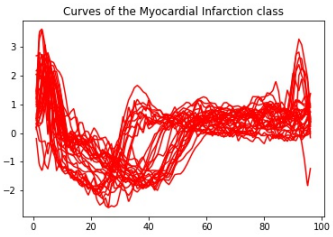
Afterwards, in the second phase, the B-spline coefficients are used as features in the input of the RF algorithm to train and validate the functional classification model.

3 An application using ECG data with a binary outcome

We apply the method described in previous section to the well-known ECG dataset [7] formatted by R. Olszewski as part of his thesis "Generalized feature extraction for structural pattern recognition in time-series data" at Carnegie Mellon University, 2001. Our objective is to create a functional model to classify ECG curves "normal" or "myocardial infarction".
Figure 1 illustrates the smoothed versions of the original signals computed using Equation (1).



(a) Functional representation of the ECG data

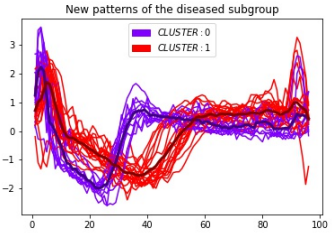


(b) Myocardial Infarction curves in the training set

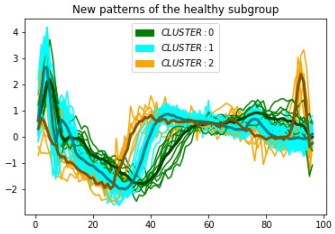
(c) Healthy curves in the training set

Fig. 1: Smoothed ECG curves in the training set.

After using the silhouette analysis with the newly defined metric for the normal and myocardial infarction classes, the most suitable number of clusters for the myocardial infarction original class is two whereas, for healthy people, we have three subgroups. Based on the latter results, the functional K-means is applied. Figure 2 shows the new patterns discovered for each subgroup in terms of functional subsets.



(a) New subgroups discovered in diseased subgroup



(b) New subgroups discovered in healthy subgroup

Fig. 2: New patterns (subgroups) discovered in the original groups of the training data.

Table 1 shows the effectiveness of the proposed two-steps functional classification approach based on the novel distance. A maximum accuracy of 87%

is obtained using RF and K-means with the new distance. Instead, using a classical supervised classification, without the two-steps procedure, we get an accuracy of 80%.

Table 1: Model performance, the combinations of Random Forest and K-means

	Random Forest			
	Without K-means	Classical K-means	K-means with the Euclidean functional distance	K-means with the new distance
Accuracy %	80	83	85	87
F1-Score	79	82	85	87
Specificity	69	64	75	83

References

1. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd Ed., Springer, New York, (2005)
2. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, (2006)
3. Zhou, Y., Sedransk, N.: Functional data analytic approach of modeling ECG T-wave shape to measure cardiovascular behavior. The Annals of Applied Statistics, pp. 1382-1402 (2009)
4. Jacques, J. and Preda, C.: Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis, 71:92–106 (2014)
5. Möller, A., Tutz, G., and Gertheiss, J.: Random forests for functional covariates. Journal of Chemometrics, 30 (12):715 – 725 (2016)
6. Huang, Q., Li, Y. and Liu, P.:Short term load forecasting based on wavelet decomposition and random forest. Proceedings of the Workshop on Smart Internet of Things. ACM. p. 2 (2017)
7. Olszewski R. Generalized feature extraction for structural pattern recognition in time-series data. Ph.d. dissertation, School of Computer Science, Carnegie Mellon University, (2001)
8. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987)

Determining shape parameters in a climate multi-model ensemble accounting for model performance and independence

Jeong-Soo Park, Yonggwan Shin, Yire Shin, and Juyoung Hong

Abstract Scientists occasionally predict the future changes in extreme climate using multi-model ensemble methods that combine predictions from individual simulation models. We employed a model weighting method that accounts for both model performance and independence (PI-weighting). In calculating the PI-weights, two shape parameters (σ_D and σ_S) should be determined, but usual perfect model test method requires a considerable computing time. To address this trouble, we suggest simple ways for selecting two shape parameters based on the chi-square statistic and the entropy, which reduce the computing time greatly. Our method is applied to 21 CMIP6 (the Coupled Model Inter-Comparison Project Phase 6) models for five climate variables over East Asia.

Key words: Climate change, Dirichlet distribution, Future prediction, Generalized extreme value distribution, Leave-one-out cross validation, Return period.

1 Introduction

Studies on the projection of future climate change have used ensembles of multiple climate simulations. Among the many Multi-model ensemble (MME) methods, model averaging is typically employed [1, 2, for example]. Model averaging is a statistical method in which unequal or equal weights are assigned to those models. Despite some arguments, the equal weighting or "model democracy" [2] has been criticized because it does not take into account the performance, uncertainty, and independency of each model in constructing an MME [3, 4, for example].

One typical unequal weighting scheme is giving more weights to those models that are more skillful and realistic for a specific process or application. This

Jeong-Soo Park, Yonggwan Shin, Yire Shin, and Juyoung Hong
Department of Mathematics and Statistics, Chonnam National University, Gwangju, Korea, e-mail: jspark@jnu.ac.kr; syg.stat@gmail.com; shinyire@daum.net; sjy0s2@naver.com

performance-based weighting method has improved the accuracy of the projections and reduced the prediction uncertainty. However, it has been reported that only a few models often exhibit extremely high weights, and most others have very low weights [5, 6]. This phenomenon may be because some models are more fit to the observations for given applications than others, and thus, receive extremely high weights in a multi-model estimate of change [7]. In addition to the performance, some researchers have considered other criteria such as model independency [3, 8, 9]. A weighting scheme that accounts for both the independence and performance simultaneously is called the PI-weighting. In this study, we employ PI-weighting to robustly quantify uncertainty in MME. In calculating the PI-weights, considering only one or two climate variables over relatively small area can lead to the overfitting problem [9, 7]. To avoid this problem, we thus consider five climate variables over the East Asia, while our focus is the annual maximum daily precipitation (AMP1) over the Korean pennisula.

In applying the PI-weighting, we have to determine two shape parameters that control the strength of the weights. One way to select the shape parameters is a leave-one-out perfect model test [9, 7], but it requires huge computing time. To overcome this computational trouble, we suggest simple ways to determine these parameters based on the entropy and p-values of the chi-square statistic.

2 Performance and independence weighting

Knutti et al.[8] argued that the growing number of models with different characteristics and considerable interdependence finally justifies abandoning a strict model democracy. As the basic idea of PI-weighting, models that agree poorly with observations for a selected set of diagnostics receive less weight, as do models that largely duplicate existing models [8]. Weights are calculated for each model based on a combination of the distance D_i (informing the performance) and the model similarity S_{ij} (informing the dependence):

$$w_i = \frac{exp(-\frac{D_i}{\sigma_D})}{1+\sum_{j\neq i}^{M} exp(-\frac{S_{ij}}{\sigma_S})}, \tag{1}$$

with the total number of model runs M and the shape parameters σ_D and σ_S. The weights are normalized such that their sum equals 1. The numerator represents the modeling skill when using a Gaussian weighting, where the weight decreases exponentially the farther away a model is from the observations. The denominator is the "effective repetition of a model" [3] and is intended to account for the model interdependency [8]. To calculate the model similarity S_{ij}, we follow a technique among several methods proposed by Sanderson et al.[10].

The shape parameters define the strength of the weighting and the relative importance of the performance and independence [7]. Large values will lead to an almost equal weighting, whereas small values will lead to aggressive (or one-sided)

weighting, giving a few models most of the weight. The shape parameters are often determined through a perfect model test (or a model-as-truth experiment) using the continuous rank probability score [7, 9]. The perfect model test picks each model from a multi-model ensemble in turn and treats it as the true representation of the climate system. This leave-one-out procedure requires huge computing time. To address this computational trouble, we consider relatively simple ways to determine the shape parameters.

2.1 Determination of σ_S

To select an appropriate value of the shape parameter σ_S for the I-weights, we consider an entropy-based approach. Denote $I_i(\sigma_S)$ as a normalized I-weight for model i and for the given σ_S. The entropy of the I-weights as a measure of uncertainty [11] from these weights is defined by the following:

$$E(\sigma_S) \;=\; -\sum_{i=1}^{M} I_i(\sigma_S)\log I_i(\sigma_S) \tag{2}$$

as a function of σ_S. When all $I_i(\sigma_S)$s are almost equal, the entropy has a high value. We thus expect the entropy to increase because σ_S has a large value.

Figure 1 presents the entropy function of σ_S computed from the data used for this study, which indicates that it is minimum at $\sigma_S = 0.4$. It is interesting to note that the entropy function increases as σ_S decreases from 0.4 to zero. This is explained by looking into the similarity measure $1+\sum_{j\neq i}^{M} exp(-\frac{S_{ij}}{\sigma_S})$. As σ_S moves toward zero, this measure converges at one for all i. Thus, s_i moves toward one, and I_i is close to $1/M$ for all i. Because we want to have a shape parameter σ_S that can differentiate the I-weights most distinctly with minimum uncertainty, the value $\sigma_S = 0.4$ minimizing the entropy is chosen in this study.

2.2 Determination of σ_D

To select an appropriate value of σ_D for the P-weights, we attempted to use the entropy criteria again, but were not fortunate enough to obtain the optimal result, as in σ_S. Thus, a technique based on the p-value of the chi-square statistic is considered in this study. Denote $P_i(\sigma_D)$ as a normalized P-weight for model i and for the given σ_D. For testing the hypothesis frame, the null hypothesis is that all weigths are equal, and the alternative hypothesis is that some weights are not equal. For the given P_i, the chi-square statistic used to test the above hypothesis is as follows:

$$\chi_0^2(\sigma_D) \;=\; \sum_{i=1}^{M} \frac{(\frac{1}{M} - P_i(\sigma_D))^2}{\frac{1}{M}}. \tag{3}$$

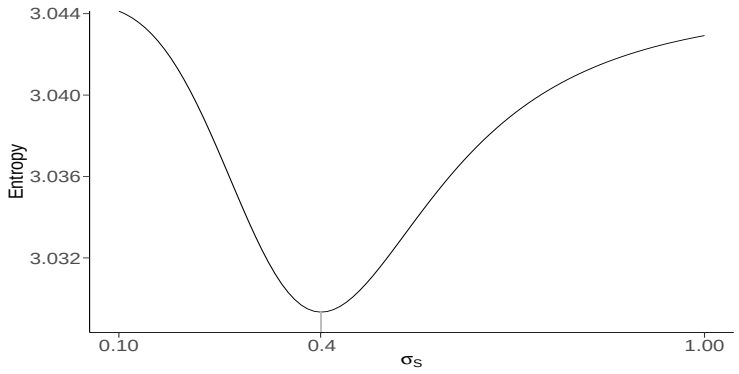


Fig. 1 Plot of the entropy as the parameter σ_S changes from 0.1 to 1.0, and the selected $\sigma_S = 0.4$.

Because we do not want to accept equal weights, σ_D should be selected to reject the null hypothesis. In addition, because we also do not want aggressive weights, a σ_D can be selected as the maximum value of σ_D in which we still reject H_0 with α level. That is, our selection is

$$\sigma_D^* = max\,\{\sigma_D : p - value\,(\sigma_D) < \alpha\}, \tag{4}$$

where $p - value\,(\sigma_D) = Pr[\chi^2 > \chi_0^2(\sigma_D)|\,H_0]$. Here, χ^2 indicates a random variable of (3) under the equal P_i weights. This selection assures the use of the least aggressive weights, and it is still statistically significantly different from the equal weights. The p-values are computed by a Monte-Carlo simulation in which random numbers of weights are generated from the Dirichlet distribution. When the parameters are all equal to 1, the Dirichlet distribution is same as the multivariate uniform distribution with values between 0 and 1, which represents the null hypothesis. We used 'MCMCpack' package [12] in R to generate the random weights satisfying H_0.

Figure 2 depicts the chi-square statistic values computed from AMP1 with some p-values as σ_D. We calculated the σ_D for each of the five climate variables, and then calculated the average from those five σ_Ds. When $\alpha = 0.05$ as is usually applied in statistics, the averaged σ_D^* from five different σ_D is 0.21.

3 Results: model weights

The normalized PI-weights are obtained using Eq.(1), with $\sigma_S = 0.4$ and $\sigma_D = 0.21$. Figure 3 demonstrates the distributions of the P-, I-, and PI-weights. The variability of the I-weights is smaller than that of the P-weights. The high P-weights of the CanESM5 and EC-Earth3-Veg models decrease in PI-weights owing to the low I-

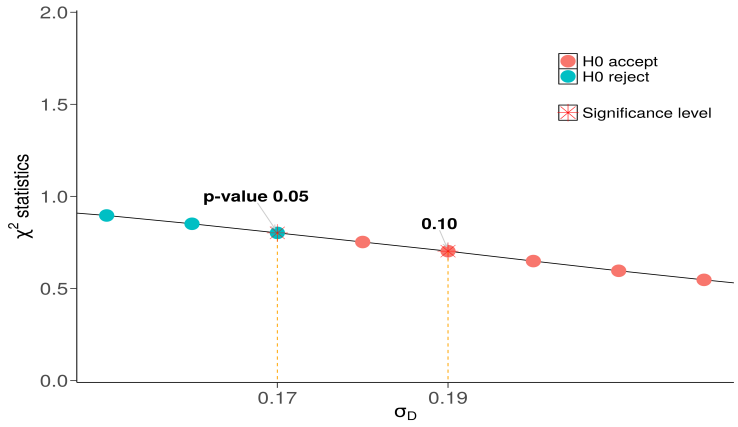


Fig. 2 Plot of the chi-square statistic values as the parameter σ_D changes, for the annual maximum daily precipitation (AMP1). The selected σ_D is 0.17 (0.19) for p-value 0.05 (0.01).

weights. The PI-weights of BCC-CSM2-MR, FGOALS-g3, and GFDL-ESM4 models increase owing to a relatively high independency. The performance is more influential to the PI-weights than the independency. Some of these observations may be changed if different σ_S and σ_D are used.

Acknowledgements This study was supported by the BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

1. Tebaldi, C., Hayhoe, K., Arblaster, J.M., Meehl, G.A.: Going to the extremes: An intercomparison of model-simulated historical and future changes in extreme events. Clim. Chang. **79**, 185–211 (2006).
2. Knutti, R.: The end of model democracy? Clim. Chang. **102**, 394–404 (2010).
3. Sanderson, B.M., Knutti, R., Caldwell, P.: A representative democracy to reduce interderpendency in a multimodel ensemble. J. Clim. **28**, 5171–5194 (2015).
4. Massoud, E.C., Espinoza, V., Guan, B., Waliser, D.E.: Global Climate Model Ensemble Approaches for Future Projections of Atmospheric Rivers. Earth's Future **7**, 1136–1151 (2019).
5. Xu, D., Ivanov, V., Kim, J., Fatichi, S.: On the use of observations in assessment of multimodel climate ensemble. Stoch. Environ. Res. Risk Assess. **33**, 1923–1937 (2019).
6. Lee, Y., Shin, Y.G., Park, J.S., Boo, K.O.: Future projections and uncertainty assessment of precipitation extremes in the Korean peninsula from the CMIP5 ensemble. Atmos Sci Lett e954 (2020).

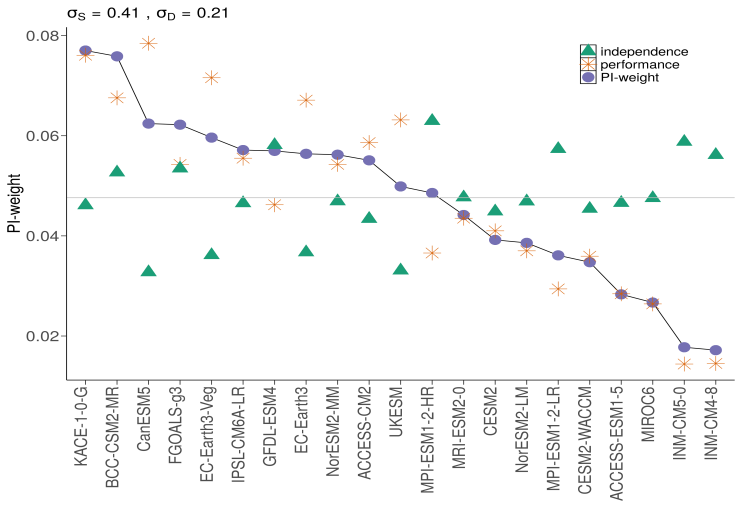


Fig. 3 Spread of the weights for 21 CMIP6 (the Coupled Model Inter-Comparison Project Phase 6) models obtained based on the performance only, the independence only, and by both the performance and independence. The weights are obtained from five climate variables over East Asia.

7. Brunner, L., Lorenz, R., Zumwald, M., Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting. Environ. Res. Lett. **14**, 124010 (2019).
8. Knutti, R., Sedlacek, J., Sanderson, B.M., Lorenz, R. et al.: A climate model projection weighting scheme accounting for performance and independence. Geophys. Res. Lett. **44**, 1909–1918 (2017).
9. Lorenz, R., Herger, N., Sedlacek, J., Eyring, V. et al.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. J. Geophys. Res. Atmos. **123**, 4509–4526 (2018).
10. Sanderson, B.M., Knutti, R., Caldwell, P.: Addressing interdependency in a multimodel ensemble by interpolation of model properties. J. Clim. **28**, 5150–5170 (2015).
11. Ross, S.: A First Course in Probability, 8th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, USA (2010).
12. Martin, A.D., Quinn, K.M., Park, J.H.: MCMCpack: Markov Chain Monte Carlo in R. J. Stat. Softw. **42**, 1–21 (2011).

Editors

Rosaria Lombardo - University of Campania "L. Vanvitelli", Italy
Ida Camminatiello - University of Campania "L. Vanvitelli", Italy
Violetta Simonacci - University of Naples "L'Orientale", Italy

PKE - Professional Knowledge Empowerment s.r.l.
Sede legale: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Sede operativa: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Ufficio Di Rappresentanza: Via Giacomo Peroni, 400 - 00131 Roma (RM)
CF / P.I. 03167830920 — www.pke.it; e-mail info@pke.it — Privacy

February 2022 PKE s.r.l.
ISBN 978-88-94593-35-8 on print
ISBN 978-88-94593-36-5 online
All rights reserved.
This work is protected by copyright law.
All rights, in particular those relating to translation, citation, reproduction in any form, to the use of illustrations, tables and the software material accompanying the radio or television broadcast, the analogue or digital recording, to publication and dissemination through the internet are reserved, even in the case of partial use. The reproduction of this work, even if partial or in digital copy, is admitted only and exclusively within the limits of the law and is subject with the authorization of the publisher. Violation of the rules involves the penalties provided for by the law.
PKE Publisher