Statistics, Technology and Data Science for Economic and Social Development

Book of short papers of the ASA Rome Conference
18 to 20 September 2024

"Measuring and Interpreting World Changes with Statistics, Data Science and AI"

Book of Short Papers of the
Sapienza University of Rome Conference

Supplement to Volume 37/2

Guest Editors:
Leonardo S. Alaimo, Fabio Crescenzi,
Luigi Fabbris, Filomena Maggino, Maurizio Vichi

Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

EDITORIAL TEAM

EDITOR IN CHIEF
- Francesco Palumbo, Università di Napoli Federico II, Naples, Italy

CO-EDITORS ON A SPECIFIC SUBJECT
- Alessandro Celegato, AICQ Centronord - Quality and techonology in production
- Adriano Decarli, Università di Milano, IRCCS /INT Foundation, Milan, Italy - Social and health studies
- Luigi Fabbris, Università di Padova, Padua, Italy - Surveys and experiments
- Vittorio Frosini, Università Cattolica del Sacro Cuore, Milan, Italy - Book review
- Antonio Giusti, Università di Firenze, Florence, Italy - Data Science
- Paolo Mariani, Università di Milano Bicocca, Milan, Italy - Social and economic analysis and forecasting

SCIENTIFIC COMMITTEE
- Thomas Aluja, UPC, Barcelona, Spain
- Paul P. Biemer, RTI and IRSS, Chicago, USA
- Jörg Blasius, Universität Bonn, Bonn, Germany
- Irene D'Epifanio, Universitat Jaume I, Castelló de la Plana, Spain
- Vincenzo Esposito Vinzi, ESSEC Paris, France
- Gabriella Grassia, Università di Napoli Federico II, Naples, Italy
- Michael J. Greenacre, UPF, Barcelona, Spain
- Salvatore Ingrassia, Università di Catania, Catania, Italy
- Ron S. Kenett, KPA Ltd. and Samuel Neaman Institute, Technion, Haifa, Israel
- Stefania Mignani, Università di Bologna Alma Mater, Bologna, Italy
- Tormod Naes, NOFIMA, Oslo, Norway
- Alessandra Petrucci, Università di Firenze, Florence, Italy
- Monica Pratesi, Università di Pisa, Pisa, Italy
- Maurizio Vichi, Sapienza Università di Roma, Rome, Italy
- Giorgio Vittadini, Università di Milano Bicocca, Milan, Italy
- Adalbert Wilhelm, Jacob University, Breimen, Germany

ASSOCIATE EDITORS
- Francesca Bassi, Università di Padova, Padua, Italy
- Bruno Bertaccini, Università di Firenze, Florence, Italy
- Matilde Bini, Università Europea, Rome, Italy
- Giovanna Boccuzzo, Università di Padova, Padua, Italy
- Maurizio Carpita, Università di Brescia, Brescia, Italy
- Giuliana Coccia, ISTAT, Rome, Italy
- Fabio Crescenzi, ISTAT, Rome, Italy
- Franca Crippa, Università di Milano Bicocca, Milan, Italy
- Corrado Crocetta, Università di Foggia, Foggia, Italy

- Cristina Davino, Università di Napoli Federico II, Naples, Italy
- Loretta Degan, Gruppo Galgano, Milan, Italy
- Tonio Di Battista, Università di Chieti-Pescara "Gabriele D'Annunzio", Pescara, Italy
- Tommaso Di Fonzo, Università di Padova, Padua, Italy
- Francesca Di Iorio, Università di Napoli Federico II, Naples, Italy
- Simone Di Zio, Università di Chieti-Pescara "Gabriele D'Annunzio", Pescara, Italy
- Filippo Domma, Università della Calabria, Rende, Italy
- Alessandra Durio, Università di Torino, Turin, Italy
- Monica Ferraroni, Università di Milano, Milan, Italy
- Giuseppe Giordano, Università di Salerno, Salerno, Italy
- Michela Gnaldi, Università di Perugia, Perugia, Italy
- Domenica Fioredistella Iezzi, Università di Roma Tor Vergata, Rome, Italy
- Michele Lalla, Università di Modena e Reggio Emilia, Modena, Italy
- Maria Cristina Martini, Università di Modena e Reggio Emilia, Modena, Italy
- Fulvia Mecatti, Università di Milano Bicocca, Milan, Italy
- Sonia Migliorati, Università di Milano Bicocca, Milan, Italy
- Michelangelo Misuraca, Università della Calabria, Rende, Italy
- Francesco Mola, Università di Cagliari, Cagliari, Italy
- Roberto Monducci, ISTAT, Rome, Italy
- Isabella Morlini, Università di Modena e Reggio Emilia, Modena, Italy
- Biagio Palumbo, Università di Napoli Federico II, Naples, Italy
- Alfonso Piscitelli, Università di Napoli Federico II, Naples, Italy
- Antonio Punzo, Università di Catania, Catania, Italy
- Silvia Salini, Università di Milano, Milan, Italy
- Luigi Salmaso, Università di Padova, Padua, Italy
- Germana Scepi, Università di Napoli Federico II, Naples, Italy
- Giorgio Tassinari, Università di Bologna Alma Mater, Bologna, Italy
- Ernesto Toma, Università di Bari, Bari, Italy
- Rosanna Verde, Università della Campania "Luigi Vanvitelli", Caserta, Italy
- Grazia Vicario, Politecnico di Torino, Turin, Italy
- Maria Prosperina Vitale, Università di Salerno, Salerno, Italy
- Susanna Zaccarin, Università di Trieste, Trieste, Italy
- Emma Zavarrone, IULM Milano, Milan, Italy

EDITORIAL MANAGER
- Domenico Vistocco, Università di Napoli Federico II, Naples, Italy

EDITORIAL STAFF
- Antonio Balzanella, Università della Campania "Luigi Vanvitelli", Caserta, Italy
- Luca Bagnato, Università Cattolica del Sacro Cuore, Milan, Italy
- Paolo Berta, Università di Milano Bicocca, Milan, Italy
- Francesca Giambona, Università di Firenze, Florence, Italy
- Rosaria Romano, Università di Napoli Federico II, Naples, Italy
- Rosaria Simone, Università di Napoli Federico II, Naples, Italy
- Maria Spano, Università di Napoli Federico II, Naples, Italy

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://www.sa-ijas.org/ojs/index.php/sa-ijas

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management



Statistica Applicata – Italian Journal of Applied Statistics (ISSN:1125-1964, E-ISSN:2038-5587) applies the Creative Commons Attribution (CC BY) license to everything we publish.

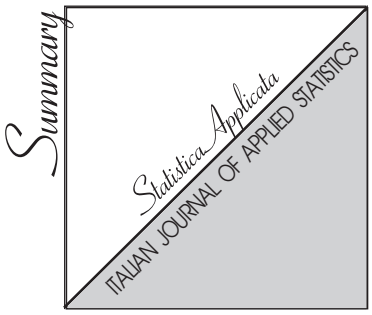
Pubblished: April 2025

© 2025Author(s)
Content license: except where otherwise noted, the present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: https://creativecommons.org/licenses/by/4.0/legalcode). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.
Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: https://creativecommons.org/publicdomain/zero/1.0/legalcode).

Published by Firenze University Press
Powered by Firenze University Press

Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.



Introduction **13**

Determinants of cooperation among hotels in Venice: A network approach **15**
Amati, V., Moretti, A., Finzi, R.M.

Modelling default risk in Italian SMEs: A spatial probit approach **21**
Amoruso, E., Bruno, E., Castellano, R., Punzo, R.

The influence of UNESCO sites on tourism demand and supply: Insights from Italian regions **25**
Antolini,F., Cesarini, S., Terragli, I.

"Inhumane": Quantitative textual analysis of a qualitative survey on gender violence among victims of international human trafficking **31**
Arcaio, M., Barbaro, F., Parroco, A.M.

Keyword-enhanced summarization from scientific literature through an integrated approach **37**
Aria, M., Cuccurullo, M., D'Aniello, L., Misuraca, M., Spano, M.

A composite indicator to measure attitude towards sustainable tourism by European citizens **43**
Bassi, F., Scioni, M.

Data quality assessment framework for composite indicator construction: An application to service delivery **49**
Bassi, F., Muhanguzi, H., Scioni, M.

Exploring art market trends: A Delphi study on collectible design **55**
Benedan, L., Codignola, F., Galeone, C., Mariani, P.

Resilience versus depression of Italian youth after the COVID-19 pandemic **61**
Bollani, L., Cascella, C., Fabbris, L.

Trust in health science and organizations in Italy during the Covid-19 pandemic **67**
Bollani, L., Fabbris. L.,

Artificial intelligence and future scenarios: Preliminary results from an application on family policies in north-eastern Italy **73**
Bolzan, M., Calleo, Y., Di Zio, S., Marozzi, M., Scioni, M.

A non-parametric test for comparative evaluations of students' performances 79
Bonnini, S., Borghesi, M., Giacalone, M., Piscopo, G.

An empirical analysis of various methods for making inferences with non-probability Samples85
Braito, L., Rocco, E.

A statistical approach for the predictive analysis of insolvency in Italy 91
Brentari, E., Ermici, G., Zanetti, A.

An innovative point of view for statistical dissemination to prevent road accidents using the 'collision matrix'.....97
Broccoli,M., Bruzzone, S.

A new algorithm for the correction and control of road accidents data based on string similarities.....103
Bruzzone, S., Valentino, L., Moretti, D.,Tavanti, M., De Blasi, F.

Advancing spatial scenarios adopting generative pre-trained transformers and text-to-image models.....109
Calleo, Y., Pilla, F., Di Zio, F.

Enhancing mobility and tourism through data analytics and generative AI 115
Cappellari, D.,Mantegar,G., Zamengo, B.

An innovative approach for the analysis of socio-economic phenomena at sub-municipal level: The Household Deprivation Study project.....121
Carbonetti, G., Biasciucci, F., Cutillo, A., Mazziotta,M., Quondamstefano, V., Tronu, D.

Volatility decomposition of traffic flow time series with complex seasonality using GARCH models.....127
Carpita, M., Metulini, R., Migliorati, M.

Economic well-being of households and attractiveness of territories: The case of Italian large municipalities.....133
Casacci, S.

An application of the partial credit model to measure the resilience of Italian young adults in the aftermath of the COVID-19 pandemic 139
Cascella, C., Bollani, L., Fabbris, L.

Data science approaches to the study of multimorbidity patterns at older ages.....145
Caselli, N., Dang, L., H., K., Rettaroli, L., Miglio, R.

Analysing fan engagement and sentiment on social media following Bologna FC's 2023/2024 Champions League qualification 151
Cavrini, G., Stavolo, A.

Text mining of tourist reviews: Sentiment analysis and trends in the Apulian tourism.....157
Centorame, D., Pino, G., Sarra, A., Nissi, E.

An exploratory analysis of the comments from the Istat's household budget survey questionnaires for designing an App 163
Cezza, E., D'Ambrosio, G., De Gaetano, L., Raimondi, E.

Combining expert knowledge and data-driven classification to support prediction and validation of census population counts.....169
Chieppa, A., Bernardini, A., Cibella, N., Solari, N.

The "group approach" versus the "inverted cascade approach" for calculating the super-dividend test of holding entities.....175
D'Achille, A., Boscaino, F.,

Migration as a diffusion process: The case of Bologna municipality.....181
D'Attoma, I., Cid, C., Tassinari, C.

Users beyond "boundaries": Recent experiences of promoting statistical literacy by Istat at the territorial level among non-regular data users 187
Da Valle, S., Osti, S., Viviano, L.C.M.

Conceptualizing and measuring educational poverty through family, school, and environmental opportunities.....193
Davino, C., De Falco, A., Fabbricatore, R., Gherghi, M., Romano, R.

The Italian NEETs after the COVID-19: What should we expect for the future? Different approaches to forecasts 199
De Luca, G., Rocca, A., Mazzocchi, P., Quintano C.

AI and geostatistics: Innovative solutions for reducing and storing CO2 205
Digrandi, A., M., Cimmino, P.

Leveraging Industry 5.0 for official business statistics innovation 211
Distefano, D., Bosso, P., Di Paolo, G., G., Papa, P.

Future orientation and psychological and social capital of Italian young adults 217
D'Uggento, A.M., Fabbris, L.

Travel and vacations of Italian senior tourists 223
D'Uggento, A.M., d'Ovidio, F., D, Marin, C., Pepe, I., Toma, E.

It's how you use the items that counts: An intelligent procedure for item selection in Item response theory.....229
Epifania,O., M., Anselmi,P., Robusto, E.

School socioeconomic status and students' achievements: Exploring territorial heterogeneity through a multilevel regression discontinuity approach 235
Fabbricatore, R., Sannino, P., Davino, C.

Capacity to aspire in high school students: Theoretical definition and measurement 241
Fabbricatore, R., Romano, R., Davino, C.

Improving online misogyny detection: The role of human annotation and explanation guided learning 247
Fontanella, L., Cucco, A., del Gobbo, E., Ignazzi, E., Fontanella, S.

The evolution of Explainable Artificial Intelligence (XAI): A preliminary systematic literature view.....253
Gnasso, A., Aria, M.

Innovative ER solutions: Using data to drive operational success 259
Grassia, M.G., Sacco, D., Paesano, D., Dagostino, F., Annachiarico, A., Di Palma, V.

Responsible AI/ML implementation in official statistics: A framework for decision-making.....265
Grippo, G., Righi, A.

Copula functions and dynamic hydrological data: The case of Scotland 269
Guerini, S.S., Spezia, L.

AI-generated sentiment in state-space models for Toyota Camry US volume sales forecasting 275
Iezzi, D., F., Monte, R., Pasquini, D.

Modeling virality in social media.....281
Iezzi, D., F., Monte,R., Pasquini, D.

Organizational flexibility measures: Remote working at the Italian National Institute of Statistics 287
Limiti, C., Sola G.

Improving information on labour demand for decision-making by integrating traditional surveys, administrative sources, and web-based data 293
Lucarelli, A., Righi, A., Rizzi, R.

Early detection of university dropouts: An experiment using video-lectures access data 299
Martini, M.C.

Statistical indicators for the analysis of social sustainability and human well-being.....305
Massari, A., Perchinunno, P., L'Abbate, S., Leogrande, D., Mongelli, D.

Health and discrimination among refugees in Germany 311
Mendola, D., Busetta, A.

Estimating overdispersion in non-transformed scores of ANVUR's department ranking procedure: A Beto-multinomial model.....317
Montanari,G., E., Doretti, M.

Design of a test for assessing teamwork and collaborative skills with competence-based knowledge space theory.....323
Morleo, F., Vitanza, A., Anselmi, P.

Efficiency of Italian courts: Awareness from spatial stochastic frontier analysis 329
Nissi, E., Sarra, A., Evangelista, A., Di Battista, T.

Campanian healthcare facilities under the microscope: Reviews, polarity, and georeferencing335
Paesano, S., Grassia, M.G., Marino, M., Mazza, R., Sacco, D., Stavolo, A.

Markov switching autoregressive models and Bayesian hydrology.....341
Paroli, R., Spezia, L.

Collecting ADS-B data for timely production of air traffic statistics 347
Pietropaoli, S., Amarone, M.

A statistical approach to measuring the impact of medication use on the lived experience of patients with inflammatory bowel disease 353
Piscitelli, A., D'Uggento, A., M., Equatore, E., Principi, M.B.

Artificial intelligence in support of accessibility: The Brain-friend device 359
Santarcangelo, V., Crisafulli, S.G., Giacalone, M., Piscopo, G., Massa, E., Vitullo, S., Scintu, D., Romano, A., Radosti, G.

Combining textual analysis and IRT to interpret the students' voice.....365
Sarra, A., Fontanella, L., Evangelista, A., Di Battista, T.

Exploring teachers' professional identity: A case study utilizing structural equation modeling and text mining in a multimodal approach 371
Sarra,A., Pentucci, M.,Fontanella. L.

Selecting sample households with high quality addresses for population census: Problems for field survey network, but bias in the estimates?.....377
Scavalli, E.

Statistical analysis of smartphone mobility data for air quality assessment 383
Shaukat, M., H., Finazzi, F.

The combination of expert knowledge and Bayesian networks to identify the relevant factors in diabetic kidney disease (DKD).....389
Slanzi, D., Poli, I., Jones, R., D.,Mayer, G.

Understanding consumer behavior in the plant-based meat market: Insights from Bayesian network analysis.....395
Slanzi, D., Lanzini, P., Checchinato, F., Colapinto, C.

Exploring childfree and childless on social media: A text analysis approach 401
Stavolo, A., Cavrini, G., Cisotto, E., Marino, M.

Mass tourism and local housing: An exploratory textual analysis.....407
Stavolo, A., Cataldo, R., Grassia, M., G., Marino, M.

Multiverse analysis: Heuristic for multi-model inference or theory of the model error?.....413
Tomaselli, V., Cantone, G., G.

From crisis to crisis: Mapping material deprivation in the EU through the 2000s.....419
Traversa, S., Ivaldi, E.

The effect of user web journey on browsing behaviour and buying attitude: The case of Sunweb.nl.....425
Urso, F., Argentino, N., Abbruzzo, A., Mohammadi, R., Pak, K., Cracolici, M., F.

Geography of business alliances and spatial network complexity: The backbone of formal collaboration agreements 429
Vurro, A., E., Bumbea, A., Giuffrida, A., Mazzitelli, A., Espa, G.

Spatial analysis of climate change awareness across countries 435
Zammarchi, G., Maranzano, P.

Future scenarios: A financial network analysis of stock prices in emerging metaverse market.....441
Zavarrone, E., Forciniti, A.

Introduction

This volume brings together a selection of papers presented at the scientific conference of the ASA (Association for Applied Statistics), held in Rome from 18 to 20 September 2024. It retains the title of the conference, "Measuring and Interpreting World Changes with Statistics, Data Science and AI." The majority of papers focuses on modeling the relationships between statistics and the emerging disciplines of data science and artificial intelligence (AI), while others address areas of applied statistics and methodological advances in official statistics.

At a time when global challenges—such as climate change, economic inequality, and technological transformation—are reshaping societies, the role of statistics and data science in interpreting and guiding these changes is more crucial than ever. This volume aims to offer a comprehensive perspective on how statistical reasoning, enhanced by modern tools from AI and data science, can contribute to understanding and addressing complex real-world issues.

All papers were anonymously reviewed by at least two independent experts from the conference's scientific committee. The published articles reflect a rigorous peer-review process and incorporate suggestions and revisions—both substantive and formal—made by the authors in response to expert feedback.

Conference presentations were grouped into thematically homogeneous sessions, covering topics such as official statistics, urban analytics, labor market dynamics, and innovations in data integration. Compared to the oral presentations, the papers in this volume have been expanded and refined, offering readers a more in-depth exploration of the subjects discussed.

Due to the self-selection of contributors and the rigorous evaluation process, the number of eligible papers has been significantly reduced. As a result, the content-based session structure could not be preserved in this publication. Instead, the articles are presented in alphabetical order by surname of the first author.

The guest editors would like to express their sincere thanks to the organizing committee and all reviewers and authors for their invaluable contributions to the success of the conference and this volume. We believe this collection will be of interest to researchers, practitioners, and policymakers engaged in quantitative analysis and evidence-based decision-making. We are confident that the insights shared in these contributions will foster further research and debate in the fields of applied statistics, data science, and artificial intelligence.

This volume will be published as a supplement to a regular issue of the international journal Statistica Applicata – Italian Journal of Applied Statistics. All articles are freely available in open access format. When citing a paper, please use the following format:

Author(s) (Year). Paper title. In L.S. Alaimo, F. Crescenzi, L. Fabbris, F. Maggino (eds.), Measuring and Interpreting World Changes with Statistics, Data Science and AI. Book of Short Papers of the ASA Rome Conference (pages). Firenze: FUP - Firenze University Press.

The guest-editors:
Leonardo S. Alaimo is tenute track researcher in Social statistics at the Department of Statistical Sciences, Sapienza University of Rome
Fabio Crescenzi is a statistical advisor and former senior researcher at Istat (Italian National Institute of Statistics), Rome
Luigi Fabbris is president of the Association for Applied Statistics and former professor of Social statistics at the University of Padua, Italy.
Filomena Maggino is full professor of Social statistics at the Department of Statistical Sciences, Sapienza University of Rome
Maurizio Vichi is a former professor of Statistics at the Department of Statistical Sciences, Sapienza University of Rome

Other members of the Scientific committee:

Fabrizio Antolini (University of Reramo)
Margaret Antonicelli (Sapienza University of Rome)
Francesca Bassi (University of Padua)
Rossella Berni (University of Florence)
Bruno Bertaccini (University of Florence)
Matilde Bini (European University of Rome)
Giovanna Boccuzzo (University of Padua)
Luigi Bollani (University of Turin)
Eugenio Brentari (University of Brescia)
Maurizio Carpita (University of Brescia)
Giulia Cavrini (University of Bolzano-Bozen)
Alessandro Celegato (AICQ-AISS, PSV Project Service and Value)
Giuliana Coccia (ASViS)
Cristina Davino (Federico II University of Naples)
Adriano Decarli (University of Milan)
Loretta Degan (Galgano Group Co., Milan)
Enrico di Bella (University of Genoa)
Angela Maria Digrandi (CNR-National Council for Research)
Simone Di Zio (G. D'Annunzio University of Chieti and Pescara)
Viviana Egidi (formerly Sapienza University of Rome)
Benito Vittorio Frosini (Sacred Heart Catholic University of Milan)
Antonio Giusti (University of Florence)
M. Gabriella Grassia (Federico II University of Naples)
Andrea Guizzardi (University of Bologna)
D. Fioredistella Iezzi (Tor Vergata University of Rome)
Paolo Mariani (University of Milan-Bicocca)
Stefania Mignani (University of Bologna)
Francesco Palumbo (Federico II University of Naples)
Chiara Parretti (Marconi University)
Alfonso Piscitelli (Federico II University of Naples)
Tommaso Pronunzio (Assirm, Ales Research)
Alessandra Righi (Istat, Rome)
Rosaria Romano (Federico II University of Naples)
Vito Santarcangelo (iInformatica, Matera)
Giuseppe Sindoni (Istat, Rome)
Giorgio Tassinari (University of Bologna, co-chair)
Nicola Tedesco (University of Cagliari)
Venera Tomaselli (University of Catania)
Maurizio Vichi (co-chair, La Sapienza University of Rome)
Domenico Vistocco (Federico II University of Naples)
Giorgio Vittadini (University of Milan-Bicocca)
Mariangela Zenga (University of Milan-Bicocca)

doi.org/10.26398/asaproc.0072

Determinants of cooperation among hotels in Venice: A network approach

Viviana Amati[a], Anna Moretti[b], Rebecca Micol Finzi[a]
[a] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy.
[b] Venice School of Management, Ca' Foscari University of Venice, Venice, Italy.

1. Introduction

Tourism is a vital economic driver, especially in Italy, where it accounted for nearly 13% of the gross domestic product in 2023 (OpenEconomics, 2023). It demands effective governance, defined as a way of coordination or interaction between destination entities operating in the third sector, to be successful and competitive (Czernek-Marszałek, 2021). In tourism destinations, cooperation between the involved stakeholders is necessary to achieve competitiveness and sustainability (Fyall and Garrod, 2020). Czernek (2013, p. 84) states that cooperation in tourism is "a form of voluntary joint actions in which autonomous stakeholders engage in an interactive process, using shared rules, norms and structures, to act and decide on issues related to tourism development in the region".

When viewed through this lens, tourism destinations can be conceived as networked systems where various entities (such as hotels and other accommodations, services, and gastronomy) collaborate and offer comparable or complementary activities. These entities are interdependent as they must work together to provide visitors with a comprehensive product—the "tourism experience". This experience encompasses all aspects of a visitor's journey, from accommodation and dining to sightseeing and entertainment, underscoring each entity's crucial role in shaping the overall experience.

Coordination and cooperation between destination organizations involve information exchange, support, training, and resource sharing. This collaboration offers benefits at both the destination and organizational levels, as highlighted by numerous studies (e.g., Elvekrok et al., 2022; Ledesma González et al., 2021). However, collaboration between very different stakeholders, such as in the tourism industry, which encompasses businesses from the hospitality sector, transportation, food and beverage, culture and creative industries, and many more, is far from being straightforward. Collaboration requires trust, reciprocal understanding, and alignment between different and sometimes conflicting interests (Czernek-Marszałek, 2019; Nunkoo, 2017). This is why, notwithstanding the well-known benefits of building networks of collaboration, they are often scantly present in tourist destinations and still have high rates of failure (Marasco et al., 2018).

In this work, we focus on Venice, a world-famous tourist destination facing the challenge of overtourism (Baldin et al., 2024). In this setting, collaboration is not only beneficial but also a crucial strategy for promoting sustainable development (Moretti et al. 2024). We conceptualize the collaboration system as a network, a system of entities (the hospitality firms in Venice) and the relationships (collaboration) among them. We apply methods typical of network analysis to explore the determinants for cooperation and the specific mechanisms that have led to establishing collaboration ties among hotels in Venice.

2. Theory background

The development of inter-firm cooperation and networks has been explored through several analytical lenses within the tourism literature, relying on well-established theories such as the transaction cost economics (Williamson, 1975), the resource-based view (Prahalad and Hamel, 1990), the resource dependency theory (Pfeffer and Salancik, 1978), and the social embeddedness

theory (Granovetter, 1973, 1985). These theories contributed to uncovering some of the explanations of network emergence (Padgett and Powell, 2012), namely why and how firms engage in collaborative relationships and invest in their development and maintenance.

Building upon these theories, reasons for collaborating between tourism firms have been analyzed mainly in terms of their benefits (e.g., in Elvekrok et al., 2022; Tribe and Liburd, 2016) and their determinants (López-Molina and Pulido-Fernández, 2024; Czernek-Marszałek, 2021).

Understanding what drives tourism firms' cooperation dynamics can be critical for understanding firm's competitiveness and the success of a destination (Baggio, 2020). Focusing on this second stream of research that serves as the main theoretical background for the present paper, we define *cooperation determinants* as those circumstances that determine firms' starting cooperation, development and results.

Determinants have been categorized into four distinct groups: preconditions of cooperation, factors crucial for the successful development of cooperation, key personal capabilities required for managing tourism cooperation, and factors conducive to cooperation effectiveness (Czernek, 2013). Between these categories, several variables have been explored, including those related to the social characteristics of the environment, geographical proximity between firms, the characteristics of the tourism governance structure, the presence of collective actors, and firms' resource endowment as a proxy of competences and experience.

Even though past studies have analyzed the determinants of cooperation, it has been raised that this problem is still insufficiently explored (Beritelli, 2011). Most of these studies adopted a qualitative approach or basic network statistics that describe different aspects of the network structure (e.g., Ledesma González et al., 2021) separately. Since the determinants act simultaneously, it is important to analyze them jointly and control their effects on one another.

In this study, we apply network models to explore the following research question: *Which factors determine the collaboration between hotels in Venice?* The network model approach allows accounting for the dependence among the collaboration ties and offers a quantitative approach which controls the effect of one factor on the tie formation for other factors.

3. Data

We gathered information about the collaboration among hospitality firms in Venice using an online questionnaire administered in 2022. The Venice School of Management of the Ca' Foscari University of Venice contacted all the hotels, B&B, hostels, and apartments in Mestre and Venice via email.

The questionnaire, designed with a recall method and a name generator, asked hospitality firms about their past collaborations and how they rated the importance of these collaborations on a scale from 1 (not important at all) to 5 (very important). Using a snowball design, we enhanced this information by conducting interviews with the hotels that did not or partially fill in the survey. We considered the hotels with four and five stars as seeds. The final sample comprises 122 hotels that declared 177 collaborations, with an average importance rating of 3.8 (ranging from 1 to 4). Most of these collaborations (41.9 %) focus on innovation and the creation of new products and services, and overbooking management (33.1 %), while smaller percentages are observed for marketing (19.8%) and training (5.2%).

We also collected information on the characteristics of the hotels. The Open Data website of the Veneto region (https://www.regione.veneto.it/web/informatica-e-e-government/open-data) provides an exhaustive data set of all the hospitality firms in Veneto with details on the facilities (e.g., private parking, restaurants, swimming pool), the quality and standards, as measured by the number of stars and "leoni" (lions: a local measure of firms' quality), and the location of the hotels (address). Starting from this data, we computed the hotel similarity with respect to the facilities using the Jaccard similarity index and the geographical distance between the hotels. Other information concerning formal relationships defined as common membership to local business associations was collected through interviews. Table 1 reports the basic descriptive statistics of the

hospitality firms' characteristics.

Table 1 – Descriptive statistics of the characteristics of the hospitality firms considered.

Stars/Leoni	n	%
1 o 2	7	0.06
3	26	21.31
4 o 5	89	72.95

	min	mean	max
Geographical distance (km)	0.4	6.5	24.95
Facility similarity	0.15	0.65	1

4. Methods

We applied valued exponential random graphs models (ERGMs) to investigate the determinants of collaboration among the hotels in Venice, with a focus on resource endowment, geographical location, and management variables.

ERGMs (Robins et al., 2010; Lusher et al. 2013) are statistical models for the analysis of cross-sectional network data, i.e., one single observation of a network. Originally developed for the analysis of binary networks, a recent extension allows the modelling of valued networks, where ties are characterized by a value indicating the intensity of the relationship (Krivitsky, 2012).

In the following, we use the convention that random variables are denoted by uppercase letters and realizations of random variables or fixed quantities are denoted by lowercase letters. We also use bold characters to indicate matrices.

We denote the observed network represented as an adjacency matrix by $\boldsymbol{X}$, where the rows and the columns are the entities, and the generic cell $\boldsymbol{x}_{ij}$ reports the intensity of the relationships between the entities *i* and *j*. We also denote the monadic and dyadic covariates by *v* and *w*, encoding the characteristics of the entities and pairs of entities, respectively. The model belongs to the family of exponential distribution and assumes that the observed network $\boldsymbol{x}$ is the outcome of the probability distribution taking the form:

$$P(\boldsymbol{X} = \boldsymbol{x}) = h(\boldsymbol{x}) \frac{exp\{\sum_k \vartheta_k s_k(\boldsymbol{x}, v, w)\}}{\kappa(\vartheta)}$$

with $h(\boldsymbol{x})$ the reference distribution, $s_k(\boldsymbol{x},v,w)$ the statistics playing the role of explanatory variables, ϑ_k parameters expressing the importance of the statistics in explaining the structure of the observed network, and $\kappa(\vartheta)$ a normalizing constant.

The statistics $s_k(\boldsymbol{x},v,w)$ are the counts of local configurations of ties representing the outcome of network mechanisms that have generated the observed network. Their choice is theory-driven, i.e. guided by substantive theories and hypotheses concerning the analyzed phenomenon.

The model specification for analyzing collaboration among the hotels in Venice includes several statistics. The first set of statistics accounts for the endogenous mechanisms describing the dependence of ties on other ties. Those statistics are *ties and intensity* defined as the number of collaboration ties and the sum of the importance values in the observed network. They capture the baseline tendency of establishing collaborations and their importance and are equivalent to the intercept in regular statistical models; *transitivity* is computed as the sum of transitive triads where ties are weighted according to their value and measures the tendency of hotels to collaborate with the collaborators of their collaborators, or in other words, the tendency for two hotels to be connected if they share a common collaborator.

The second set of statistics relates to the exogenous mechanisms describing the dependence of ties on the hotel attributes. Those statistics are the *similarity* defined as the sum of the values of ties between hotels having a similar characteristic and modelling the tendency of hotels to collaborate with hotels that have similar characteristics; the *activity* defined as the sum of the values of ties weighted by the attribute of the hotel and modelling how the hotel characteristics influence the establishing of collaboration ties.

The parameters are estimated using algorithms based on Monte Carlo Markov chains. A positive (or negative) value implies that a larger (or smaller) number of configurations of type k is observed in $\boldsymbol{x}$ than expected by chance, thereby providing evidence for (or against) the corresponding mechanism.

5. Results

Table 2 presents the estimate of the valued ERGMs, with all parameters showing significant results at a 5% level but the activity parameters.

Table 2 – Valued ERGMs for the importance of collaboration among hotels. Parameters estimates (Est.), standard errors (s.e.), and significance (Sign.)

Statistic	Est.	s.e.	Sign.
Endogenous effects			
Ties	-6.16	0.3	***
Intensity	-0.61	0.31	*
Transitivity	0.52	0.09	***
Exogenous effects			
Stars activity	0.11	0.18	
Stars similarity (absolute difference)	-0.19	0.09	*
Facility activity	0.14	0.16	
Facility similarity	0.16	0.07	*
Same business association	0.49	0.13	***
Geographical distance (logarithm)	-0.07	0.03	*

***p < 0.001; **p < 0.01; *p < 0.05.

We start by commenting on the endogenous effects. The negative coefficients of the *ties* and *intensity* statistics indicate that the network is sparse, with only a few collaboration links. This finding aligns with the network's low density (observing only 3% of all possible collaboration ties). The positive value of the *transitivity* statistic suggests that collaboration ties tend to cluster and that hotels sharing a common collaborator are more likely to cooperate.

Now, we consider the exogenous effects. The positive coefficient of the *stars activity* is not significant, and thus, hotels with high and low quality show the same tendency to entertain collaboration. The negative parameter of the *stars similarity*, measured by the absolute difference between the stars of the hotels, indicates that hotels tend to collaborate with those of similar quality. Similarly, the positive parameter of the *activity* related to the number of offered facilities is not significant, suggesting that there is no difference in the collaboration tendency of hotels with high and low numbers of provided services. The positive coefficient of the facility *similarity* indicates a tendency for hotels to collaborate with hotels that offer the same services.

Finally, the positive coefficient related to being a member of the same business association underscores the significant influence of formal relationships on collaboration. In contrast, the negative coefficient of the geographical distance suggests that nearby hotels are more likely to collaborate, further emphasizing the influence of proximity on collaboration.

5. Conclusions and discussions

In this work, we investigated the determinants of collaboration between hospitality firms in Venice while accounting for the importance assigned to these relationships. The analysis indicates that transitivity plays a crucial role in shaping collaboration patterns in the hotel industry. Specifically, hotels tend to collaborate in small groups, as suggested by the positive and significant parameter related to the transitivity mechanism.

Homophily, that is the tendency of entities to associate and bond with similar others, is also an essential driver of collaboration. The model provides evidence for the tendency of hotels to cooperate with hotels with the same quality and similar services, a manifestation of homophily in the context of the hotel industry. This result might be justified, in part, by the fact that cooperation might take place to deal with overbooking, namely the mechanism that hotels overbooked use to redirect their guests to alternative accommodation solutions — that must be of the same quality level. From another point of view, hotels similar to each other often share the same business approach (e.g., managerial *vs.* family business), and thus, they share a common "language" that facilitates trust, development and collaboration (Nunkoo, 2017).

Membership in business associations, which often provide a platform for networking and collaboration, is also a determinant of collaboration. This finding might highlight the importance of these associations in fostering collaboration in the hotel industry, as they can facilitate the formation of collaborative ties among member hotels, as discussed by previous work (Moretti et al., 2024).

Finally, the urban and territorial morphology of Venice explains the significant role of the geographical distance in shaping the collaboration network. Venice has a very small historical center, geographically separated from the mainland by a bridge. Its structure, articulated on over a hundred islands, emphasizes the relevance of being "neighbors" only when the geographical proximity is very high: a distance that in other cities could be considered negligible, in Venice, can be relevant. Overbooking mechanisms, as described above, can reinforce relationships only between firms located very close to each other.

This work is a first attempt to analyze the collaboration of hotels in Venice and serves as a pilot study for more extensive data collection and analysis. It is important to note that the study has some limitations. The first is the representativeness of our sample: hotels with one or two stars are underrepresented due to the sampling design and the response rate during the data collection. The second is the presence of missing data. A non-response generates missing ties that could modify the structure of the network and impact on inference. Thus, the results have only a descriptive value and our findings cannot be generalized.

While this work analyzes only a small piece of the collaboration puzzle between hospitality firms in Venice, it paves the way for future research on determinants of network collaboration and provides a framework that can be applied to any other context to understand collaboration between hospitality firms.

References

Baggio, R. (2020). Tourism destinations: A universality conjecture based on network science. *Annals of Tourism Research,* **82**, 102929.

Baldin, A., Bertocchi, D., Camatti, N. (2024). Crowding-in and (temporary) crowding-out in Venice. The effect of cultural events on residents. *Tourism Economics*, https://doi.org/10.1177/13548166241233773.

Beritelli, P. (2011). Cooperation among prominent actors in a tourist destination. *Annals of Tourism Research,* **38**(2), pp. 607-629.

Czernek, K. (2013). Determinants of cooperation in a tourist region. *Annals of Tourism Research,* **40**, pp. 83-104.

Czernek-Marszałek, K. (2019). Applying mixed methods in social network research – The case of cooperation in a Polish tourist destination. *Journal of Destination Marketing & Management,*

11, pp. 40-52.
Czernek-Marszałek, K. (2021). The sources and components of social embeddedness as determinants of business cooperation in a tourist destination. *Journal of Destination Marketing & Management*, **19**, 100534.
Elvekrok, I., Veflen, N., Scholderer, J., Sørensen, B.T. (2022). Effects of network relations on destination development and business results. *Tourism Management*, **88**, 104402.
Fyall, A., Garrod, B. (2020). Destination management: a perspective article. *Tourism Review,* **75**(1), pp. 165-169.
Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology,* **78**(6), pp. 1360-1380.
Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology,* **91**(3), pp. 481-510.
Krivitsky, P.N. (2012). Exponential-family random graph models for valued networks. *Electronic journal of statistics*, **6**, 1100.
Ledesma González, O., Merinero-Rodríguez, R., Pulido-Fernández, J.I. (2021). Tourist destination development and social network analysis: What does degree centrality contribute?. *International Journal of Tourism Research*, **23**(4), pp. 652-666.
López-Molina, D., Pulido-Fernández, J.I. (2024). Analysis of the interaction and collaboration of the actors of a tourist destination from a network perspective. *Current issues in tourism,* **27**(10), 1572-1589.
Lusher, D., Koskinen, J., Robins, G. (2013). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press.
Marasco, A., De Martino, M., Magnotti, F., Morvillo, A. (2018). Collaborative innovation in tourism and hospitality: A systematic review of the literature. *International Journal of Contemporary Hospitality Management*, **30(5)**. DOI:10.1108/IJCHM-01-2018-0043
Moretti, A., Martini Barzolai, M., Cutugno, M. (2024). Ambidexterity of hospitality trade associations: A social capital perspective for destination coordination and cooperation dynamics. *Current Issues in Tourism*. DOI:10.1080/13683500.2024.2414929
Nunkoo, R. (2017). Governance and sustainable tourism: What is the role of trust, power and social capital? *Journal of Destination Marketing & Management,* **6**(4), 277-285.
OpenEconomics (2023). Il turismo in Italia. Impatto della spesa turistica sull'economia. https://www.openeconomics.eu/wp-content/uploads/2023/12/OE_ImpattoTurismo.pdf
Padgett, J.F., Powell, W.W. (2012). The problem of emergence. In J.F. Padgett, W.W. Powell (Eds.), *The Emergence of Organizations and Markets* (pp. 1-29). Princeton: Princeton University Press.
Pfeffer, J., Salancik, G.R. (1978). *The External Control of Organizations: A Resource Dependence Approach.* NY: Harper and Row Publishers.
Prahalad, C., Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review,* ***68***(3), pp. 295-336.
Robins, G., Pattison, P., Kalish, Y., Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, **29**(2), pp. 173-191.
Tribe, J., Liburd, J.J. (2016). The tourism knowledge system. *Annals of Tourism Research,* **57**, pp. 44-61.
Williamson, O.E. (1975). Markets and hierarchies: analysis and antitrust implications: A study in the economics of internal organization. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*. Free Press.

doi.org/10.26398/asaproc.0073

Modelling default risk in Italian SMEs: A spatial probit approach

Emanuele Amoruso[a], Emma Bruno[b], Rosalia Castellano[b], Gennaro Punzo[a]

[a] Department of Economic and Legal Studies, University of Naples Parthenope, Naples, Italy.
[b] Department of Management and Quantitative Studies, University of Naples Parthenope, Naples, Italy.

1. Introduction

Default risk analysis has increasingly become crucial for banks and rating agencies (Altman et al., 2020), as they require analytical tools to effectively evaluate loan approvals and creditworthiness, especially for small and medium-sized enterprises (SMEs) (Modina et al. 2023; Ciampi et al., 2021). Unlike large companies, which undergo financial statement audits, lending to SMEs is often riskier due to their less transparent financial disclosures and non-audited financial statements (Calabrese et al., 2019). Moreover, small firms face greater challenges in accessing credit and obtaining financial resources and encounter stricter borrowing requirements from financial institutions compared to larger firms (Wang and Guedes, 2024).

In the Italian economic context, SMEs form the backbone of the business landscape, representing nearly all Italian companies. SMEs employ approximately 76.5% of the workforce and contribute around 65% to the creation of added value at factor cost (Istat, 2023). Analysing their financial stability is essential for understanding overall economic health and making informed loan approvals by financial institutions. Therefore, constant and rapid monitoring of both internal risk factors and firms' financial performance is needed to develop predictive models for insolvency or acquisition likelihoods (Altman et al., 2023).

Alongside internal factors, it is reasonable to assume that a firm's location provides additional information, particularly in assessing the default risk of SMEs, which are presumed to be more influenced by local economic conditions (Fernandes and Artes, 2015). Recent research (Calabrese, 2023; Medina-Olivares et al., 2022; Chen and Lee, 2020) underscores the necessity of considering interdependences among businesses in credit scoring models (Calabrese et al., 2019), as the default probability of firms can be mutually affected by the performance of neighbouring businesses.

This research aims to evaluate how spatial dependencies alongside firms' financial indicators contribute to assessing default risk using spatial probit models on 2022 firm-level data of Italian SMEs. These models account for endogenous and exogenous interaction effects by integrating the spatial lag in the target variable and predictors. This approach intersects with spatial finance, an emerging field that integrates spatial data into financial theory and practice (Caldecott et al., 2022). Spatial finance leads the forefront of future financial analysis by spatially modelling information markets, risk assessment, valuation techniques, and investment opportunities, paving the way for innovative financial decision-making and risk management approaches.

2. Methods and data

The data were sourced from the AIDA database, which includes economic and financial performance indicators, as well as demographic and commercial information for all active and bankrupt joint-stock Italian companies with a history spanning up to ten years. The AIDA database is georeferenced, providing location-based information for each company. Focusing on SMEs, a sample of 9,292 firms with fewer than 250 employees and annual turnover or balance sheet totals not exceeding 50 million euros or 43 million euros, respectively, were extracted in the latest

available year, 2022. It serves as a representative sample for evaluating credit ratings, assessing creditworthiness, and constructing predictive models for insolvency or acquisition likelihoods.

To develop a supervised learning model, it is crucial to establish a target variable for classifying firms as solvent or insolvent. This variable was based on the legal status of SMEs, categorising them as: active in insolvency status, bankrupt, ceased, closed due to bankruptcy, and active but with annual total debts exceeding 500,000 euros. These firms were coded as '1', while all other active solvent firms were coded as '0'. The choice of covariates was informed by relevant literature (Altman et al., 2020), while accounting for constraints related to data availability and missing data issues: i) EBITDA to sales ratio; ii) Banks liabilities to turnover ratio; iii) Debts to EBITDA ratio; iv) Debts to equity ratio; v) Liquidity index; vi) Debt ratio; vii) Short-term investment index; viii) Employee costs; and ix) ROI.

To simultaneously explore the predictors of default probability among SMEs and the role of spatial proximity between firms, two probit spatial models were proposed (Lacombe and LeSage, 2018; McMillen, 1992): the probit spatial autoregressive (PSAR) model, including spatial lag only in the target variable:

$$\Phi^{-1}\big(P_r(Y_i = 1)\big) = \rho WY + X\beta + \epsilon \quad with \;\; \epsilon \sim N(0, \sigma^2 I) \quad (1)$$

and the probit spatial Durbin model (PSDM), which additionally includes spatial lag among predictors:

$$\Phi^{-1}\big(P_r(Y_i = 1)\big) = \rho WY + X\beta + \theta WX + \epsilon \quad with \;\; \epsilon \sim N(0, \sigma^2 I) \qquad (2)$$

where P_r denotes probability, Y is the $n\ x\ 1$ vector of the binary target variable with a Bernoulli distribution $Y \sim Be\big(\Phi(X^T\beta)\big)$, X is the $p\ x\ n$ matrix of covariates, β is the $p\ x\ 1$ vector of unknown parameters, Φ denotes the cumulative distribution function of the standard normal distribution, ρ is the autoregressive parameter, W is the spatial weight matrix, and ε is the error term; θ in equation 2 is the $p\ x\ 1$ vector of unknown parameters of lagged covariates.

The parameter estimation for both probit spatial models was based on maximum likelihood. A *k*-nearest neighbours (*knn*) spatial weight matrix was used with different neighbour counts (5, 10, 15, 20, and 25) to determine the model's optimal specification. The choice of *k* was informed by classifier accuracy indicators, such as sensitivity and specificity, as well as the model's ability to capture spatial autocorrelation.

3. Results and discussion

This section discusses the results of the probit model without spatial effects, the probit spatial autoregressive (PSAR), and the probit spatial Durbin model (PSDM) for 2022 (Table 1). Probit spatial models analyse the causal relationship between firms' status (insolvent or solvent) and economic and financial predictors selected via the LASSO method. LASSO shrinks the estimated regression coefficients and sets those with low explanatory power to zero, retaining only those that ensure the best model fit, thereby avoiding multicollinearity. The model's complexity is determined by both the original predictors selected by LASSO and the spatially lagged predictors, including the spatial lag of the target variable.

Both the PSAR and PSDM with *k*=15 show the best performance compared to other *k* values. Accuracy further improves with PSDM, showing a 24.5% increase compared to the probit model without spatial effects. Sensitivity (the ability to correctly predict defaulting firms – improved by 15.8%, while specificity (the ability to correctly predict solvent firms) improved by 48.1%. Moreover, PSDM specification has nearly eliminated spatial autocorrelation in the residuals, achieving an 87.7% reduction and resulting in negligible Moran's I residuals. Further supporting evidence comes from the ROC curves, where the PSDM specification shows an area of 0.984 compared to 0.816 for the probit model without spatial effects.

Table 2. Probit without spatial effects, PSAR, and PSDM estimates for k=15

	Probit without spatial effects	**PSAR**	**PSDM**
Intercept	-0.286***	-1.843***	-1.157***
Rho	–	0.739***	0.686***
Liquidity index	-0.178***	-0.474***	-0.579***
Banks liabilities to turnover ratio	0.012***	0.011***	0.047***
Employee costs	0.002***	0.005***	0.006***
Debts to EBITDA ratio	0.034***	0.101***	0.109***
W*Liquidity index	–	–	-0.423***
W* Banks liabilities to turnover ratio	–	–	-0.088***
W*Employee costs	–	–	0.001***
W*Debts to EBITDA ratio	–	–	0.124***
Accuracy	0.746	0.907	0.929
Sensitivity	0.819	0.932	0.948
Specificity	0.603	0.858	0.893
Moran' s I residuals	0.381***	0.094	0.047

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

The positive spatial autocorrelation coefficient shows spatial dependency among neighbouring firms. Such dependency arises because firms located near each other may share common risks or be influenced by local factors, such as regional economic policies, market conditions, or shared infrastructure. Consequently, these shared influences likely lead to similar economic and financial characteristics, resulting in comparable default risks.

The results highlight several factors influencing firms' default risk and reveal how the economic and financial performance of both firms and their neighbours can affect the probability of default. Specifically, a higher ratio of banks' liabilities to turnover increases default risk, indicating that firms with a greater exposure to bank debts are more vulnerable to default. However, this effect is partially mediated by the ratio of banks' liabilities on turnover of neighbouring firms, meaning that firms' default risk may be mitigated by their operational context and the debt situation of nearby businesses. Higher employee costs increase default risk, suggesting that firms burdened with higher workforce-related expenses may encounter heightened financial strain. In contrast to the spatial effect of banks' liabilities to turnover, high employee costs among neighbouring firms may intensify default risk. A higher debt-to-EBITDA ratio indicates higher default risk, reflecting firms with substantial debt relative to their earnings facing increased vulnerability to default. The effect of neighbouring firms' debt-to-EBITDA ratio reinforces the positive impact on default risk, suggesting systemic financial challenges across adjacent firms. On the other side, a higher liquidity index decreases default risk, indicating that firms with stronger liquidity status are less likely to default due to their greater ability to meet financial obligations. Similarly, the liquidity status of neighbouring firms reinforces the dampening effect on default risk.

4. Conclusion

The results drawn from this research highlight both the data modelling outcomes and the practical implications. The study demonstrates that both the estimated spatial probit models outperform traditional probit models without spatial effects, providing more efficient results and

proving to be effective tools for assessing default risk. The findings underscore the importance of considering spatial dependencies in credit risk assessment to enable financial institutions to make more informed lending, as well as implications for policy and corporate strategy in managing default risk among SMEs. Factors such as banks' liabilities to turnover, employee costs, debt-to-EBITDA ratios, and liquidity indexes play crucial roles in shaping default probabilities among neighbouring firms.

Effective regional economic policies should focus on fostering financial networks, targeted monitoring, and intervention strategies tailored to economic diversification. Such measures can enhance resilience against economic downturns and mitigate systemic risks across neighbouring firms. Strategies centred on collaborative risk management, optimising labour costs, and careful liquidity management are crucial for SMEs. These approaches not only strengthen firms' financial health but also contribute to collective resilience within local business ecosystems. Financial institutions are encouraged to integrate spatial context into risk assessments and offer diversified financial solutions tailored to varying risk profiles among SMEs.

References

Altman, E.I., Balzano, M., Giannozzi, A., Srhoj, S. (2023). Revisiting SME default predictors: The omega score. *Journal of Small Business Management,* **61**(6), pp. 2383-2417.

Altman, E.I., Esentato, M., Sabato, G. (2020). Assessing the credit worthiness of Italian SMEs and mini-bond issuers. *Global Finance Journal*, **43**, 100450.

Calabrese, R. (2023). Contagion effects of UK small business failures: A spatial hierarchical autoregressive model for binary data. *European Journal of Operational Research*, ***305***(2), pp. 989-997.

Calabrese, R., Andreeva, G., Ansell, J. (2019). "Birds of a feather" fail together: Exploring the nature of dependency in SME defaults. *Risk Analysis*, **39**(1), pp. 71-84.

Caldecott, B., McCarten, M., Christiaen, C., Hickey, C. (2022). Spatial finance: Practical and theoretical contributions to financial analysis. *Journal of Sustainable Finance & Investment*, pp. 1-17.

Chen, T. H., Lee, C. C. (2020). Spatial analysis of liquidity risk in China. *The North American Journal of Economics and Finance*, **54**, 101047.

Ciampi, F., Giannozzi, A., Marzi, G., Altman, E.I. (2021). Rethinking SME default prediction: A systematic literature review and future perspectives. *Scientometrics,* **126**, pp. 2141-2188.

Fernandes, G.B., Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research,* **249**(2), pp. 517-524.

Istat (2023). *Annuario Statistico Italiano*. Istat.

Lacombe, D. J., LeSage, J. P. (2018). Use and interpretation of spatial autoregressive probit models. *The Annals of Regional Science*, **60**, pp. 1-24.

McMillen, D.P. (1992). Probit with spatial autocorrelation. *Journal of Regional Science*, **32**(3), pp. 335-348.

Medina-Olivares, V., Calabrese, R., Dong, Y., Shi, B. (2022). Spatial dependence in microfinance credit default. *International Journal of Forecasting*, ***38***(3), pp. 1071-1085.

Modina, M., Pietrovito, F., Gallucci, C., Formisano, V. (2023). Predicting SMEs' default risk: Evidence from bank-firm relationship data. *The Quarterly Review of Economics and Finance,* **89**, pp. 254-268.

Wang, W., Guedes, M.J. (2024). Firm failure prediction for small and medium-sized enterprises and new ventures. *Review of Managerial Science,* pp. 1-34.

doi.org/10.26398/asaproc.0074

The influence of UNESCO sites on tourism demand and supply: Insights from Italian regions

Fabrizio Antolini, Samuele Cesarini, Ivan Terraglia
Department of Business Communication, University of Teramo, Teramo, Italy

1. Introduction

The inscription of sites on the UNESCO World Heritage List is widely recognized as a powerful tool for promoting the conservation of cultural and natural heritage. However, in addition to cultural enhancement and protection, UNESCO designation has a significant impact on tourism demand and supply in the regions concerned. Numerous empirical studies show that inclusion on the World Heritage List acts as a catalyst for tourism, increasing visitor numbers and contributing significantly to the local economy (Patuelli et al., 2013; Guccio et al., 2017).

Despite the economic benefits associated with increased tourism flows, the effectiveness of UNESCO designations in promoting tourism is not uniform and depends on several environmental factors. The nature of the sites themselves, which may be cultural or natural heritage sites, and the diversity of regional contexts profoundly influence tourism attractiveness and the dynamics of supply and demand (Huang et al., 2012; Moscardo, 1996). Furthermore, the sustainable management of sites is crucial to maintain a balance between heritage conservation and tourism development. There is a need to address the challenges of increasing tourist flows while ensuring the preservation of the unique characteristics that make these sites worthy of protection (King and Halpenny, 2014; Landorf, 2009). The importance of careful and sustainable management is further underlined by the need to integrate strategies that consider local specificities and enhance the distinctive elements that characterise heritage value. Regional variability and the nature of sites require customised approaches to maximise tourism benefits without compromising long-term sustainability (Şahin Körmeçli, 2024).

The objective of the study is, therefore, to analyse the relationship of tourism demand and supply in the Italian regions, identified respectively in the number of accommodation facilities and tourist nights spent, with the presence of UNESCO heritage sites on the territory, and to understand whether the areas of highest density of the two indicators coincide with the places where UNESCO sites are located.

After this brief introduction, the discussion proceeds with a description of the data used in the analysis and an explanation of the methodology used to relate the demand and supply of tourism in the regions to the presence of UNESCO heritage sites on the territory. Some results will then be presented, in particular those of the region of Sicily. At the end of the discussion, there will be a discussion of the results, the limits of the research and future implications.

2. Data and methodology

The study, therefore, takes into consideration two indicators used as proxies for tourism demand and supply, the number of municipal accommodation facilities and municipal tourism nights spent respectively, and analyses their distribution over the regional territory concerning the presence of UNESCO heritages and their nature. It was decided to use only the 59 material heritages, as they are more easily geo-localised at one point on the regional territory.

The two indicators mentioned are derived from ISTAT surveys, the first - the number of accommodation establishments - is from the 'Capacità degli esercizi ricettivi' survey, which surveys, at the level of the individual municipality, the number of establishments, beds, rooms and

bathrooms for hotel establishments; establishments and beds for other establishments (Istat, 2023). While the second one - tourism nights spent - is to be found in the survey 'Movimento dei clienti negli esercizi ricettivi' (Istat, 2023). For both, data were collected for each Italian municipality and were inserted into the cartograms of the respective regions, aligning the name of the municipality with the identification code so that the indicators were included in the area occupied by the corresponding municipality. The municipal-level territory was chosen for the analysis to capture as more as possible local tendencies in the indicators identified as proxies of tourism's demand and supply (Netrdová et al., 2020).

At the same time, on the same cartograms, all UNESCO World Heritage sites were entered by locating them using the geographical coordinates (latitude and longitude) provided by UNESCO's official list (UNESCO, 2024), extracting and numbering them manually according to the order of extraction.

To relate the presence of UNESCO heritage sites in the territory to the indicators mentioned above, it was considered to calculate density estimates of tourism supply and demand using Kernel Density Estimation (KDE) (Equation 1). The latter is a non-parametric interpolation technique used to identify specific areas (Antolini and Truglia, 2023). KDE uses a three-dimensional moving function that weighs events within its sphere of influence according to the distance from the point of intensity estimation (Bailey and Gatrell, 1995). The formula shown represents the intensity of the distribution of point events observed at a specific location, using a Gaussian kernel distribution and a smoothing parameter that considers the spatial variability of the observations in the region:

$$\lambda_s^* = \Sigma_{i=1}^{n} 1/t^2 * g(s-s_i/t) \quad \textbf{(1)}$$

Where:

- λ_s^* is the estimated intensity of the distribution of point events observed at the location s.
- s is the point of interest in space for which the density is being estimated.
- si are the points where events actually occur.
- n is the total number of observed events.
- g(·) is the kernel function, a weight function that determines how events affect the density estimate in s, in this case a Gaussian kernel function is used, with events closer to s receiving larger weights
- t is the smoothing parameter, which controls the smoothing effect and considers the spatial variability of the observations in the region.

3. Results

With the application of the previously illustrated methodology, density estimates were therefore calculated for each Italian region of tourism nights spent and municipal accommodation facilities. The results of only one region, Sicily, will be presented and commented on below, relating them to the presence of UNESCO heritage sites on the territory. The choice of this region was due to the presence of the largest number of its own UNESCO heritage sites, namely 7 (a record shared with the region of Tuscany). They are listed below (**Table 1**):

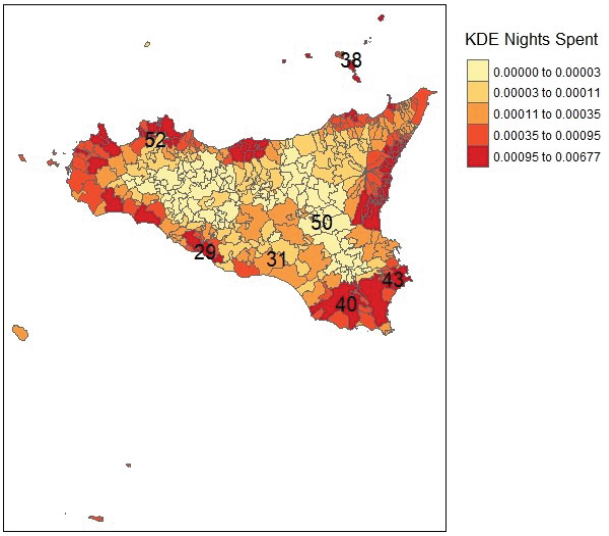
Table 1. UNESCO World Heritage Sites in Sicily

Site	**ID**
Agrigento Archaeological Area	29
Roman Villa del Casale	31
Aeolian Islands	38
Late Baroque towns of Val di Noto	40

Syracuse and the Necropolis of Pantalica	43
Mount Etna	50
Arab-Norman Palermo and the cathedrals of Cefalù and Monreale	52

Through their coordinates (latitude and longitude), they were placed on maps with the help of the R programming software, so that their presence or absence could be contextualised in areas of high or low density of the indicators taken into consideration and with the ultimate aim of understanding the real attractive potential of UNESCO heritage sites and possible areas of intervention in less virtuous territories.

Figure 1. Kernel density estimation of Sicily nights spent with UNESCO Heritages



The map (**Figure 1**), therefore, represents the estimation of the density of municipal tourism nights spent in Sicily using the Kernel Density Estimation (KDE) technique. The legend indicates different density bands, with colours ranging from light yellow (low density) to dark red (high density).

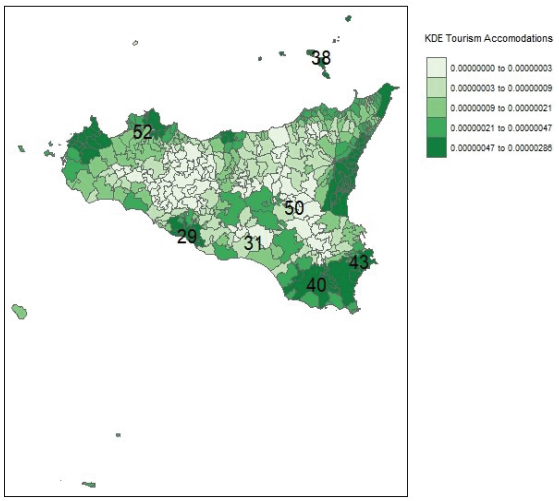


Figure 2. Kernel density estimation of Sicily tourism accommodations with UNESCO heritages

Similarly to the previous one, **Figure 2** represents the density estimates of the indicator used

as a proxy for tourism supply, the accommodation facilities in the municipalities of Sicily. Again, the legend indicates different density bands, with colors ranging from light green (low density) to dark green (high density).

As can be seen from the maps, the areas with the highest density, for both indicators, are distributed along the coastline and mainly around the region's largest cities. In most cases, these also coincide with areas that include UNESCO heritage sites, as in the case of the archaeological area of Agrigento, the Late Baroque towns of Val di Noto, Syracuse and the necropolis of Pantalica, and Arab-Norman Palermo and the cathedrals of Cefalù and Monreale. The Aeolian Islands also show a very high density for both indicators, unlike the other two UNESCO heritage sites, namely the Roman Villa del Casale and Mount Etna.

4. Discussion, conclusions and future implications

There is a clear relationship between areas with a high density of tourism nights spent and accommodation facilities and the presence of UNESCO sites in Sicily. These sites act as powerful attractors for visitors and tourists, requiring an adequate accommodation infrastructure to welcome them. Areas with UNESCO sites often coincide with the darker coloured areas in the maps of nights spent and accommodation facilities, indicating that cultural heritage significantly influences both the influx of tourists and the distribution of tourism infrastructure on the territory. This is not always the case when it comes to natural UNESCO heritages, e.g. Mount Etna. In this case, for clear problems due to the geographic conformation of the territory, it is not able to register high numbers of nights spent and accommodation facilities in the municipalities that include it. A different argument could be made for the Aeolian Islands, a natural UNESCO heritage site, which, however, has a high level of density in terms of both supply and demand.

Thus, the nature of the UNESCO heritage site itself is likely to influence the tourism supply and demand of the area in which it is located. As can be deduced from the results of this analysis, if the site, whether cultural or natural, is set in a context of cultural or even other types of tourist attractions (e.g. seaside), it can contribute to high performance. Conversely, it is possible that some sites, if located in territories that are difficult to access or even in protected natural areas, and therefore legally restricted, may emit a poor performance output, such as the case of Mount Etna.

This study is limited to the analysis of tourism supply and demand using tourism nights spent and accommodation facilities as proxies, thus maintaining a very general level of tourism performance. In the future, the analysis could be extended to include more indicators, such as cultural tourism supply, in order to achieve more specific levels of definition of the tourism landscape in which these attractions are embedded.

Also at the methodological level, the study focuses on analysis by means of kernel density estimation using only one type among the possible ones for the kernel function, in this case the Gaussian. Another future implication could be to compare kernels with different functions and then use the most suitable one.

Statements and Declarations

The authors did not receive support from any organization for the submitted work.

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Research data for this article

The scripts used in R can be requested as needed.

References

Antolini, F., Truglia, F. G. (2023). Using farmhouse and food to enforce a tourism sustainable development model: empirical evidence from Italy. *National Accounting Review*, 5(2), pp. 159-173.

Bailey, T. C., Gatrell, A. C. (1995). *Interactive Spatial Data Analysis (Vol. 413, No. 8)*. Essex: Longman Scientific & Technical.

Guccio, C., Levi Sacerdotti, S., Rizzo, I. (2017). An empirical investigation of cultural travellers' preferences and behaviours in a destination with mixed environmental features. *Enhancing Participation in the Arts in the EU: Challenges and Methods*, 249-265. https://doi.org/10.1007/978-3-319-09096-2_17.

Huang, C. H., Tsaur, J. R., Yang, C. H. (2012). Does world heritage list really induce more tourists? Evidence from Macau. *Tourism Management*, 33(6), pp. 1450–1457. https://doi.org/10.1016/j.tourman.2012.01.014.

Istat. (2023). *Capacità degli esercizi ricettivi.* https://siqual.istat.it/SIQual/visualizza.do?id=5000158&language=IT.

Istat. (2023). *Movimento dei clienti negli esercizi ricettivi.* https://www.istat.it/informazioni-sulla-rilevazione/movimento-dei-clienti-negli-esercizi-ricettivi-anno-2010/.

King, L. M., Halpenny, E. A. (2014). Communicating the World Heritage brand: Visitor awareness of UNESCO's World Heritage symbol and the implications for sites, stakeholders and sustainable management. *Journal of Sustainable Tourism*, 22(5), pp. 768–786. https://doi.org/10.1080/09669582.2013.864660.

Landorf, C. (2009). Managing for sustainable tourism: A review of six cultural world heritage sites. *Journal of Sustainable Tourism*, 17(1), pp. 53–70. https://doi.org/10.1080/09669580802159719.

Moscardo, G. (1996). Mindful visitors: Heritage and tourism. *Annals of Tourism Research*, 23(2), pp. 376–397. https://doi.org/10.1016/0160-7383(95)00068-2.

Netrdová, P., Nosek, V., Hurbánek, P. (2020). Using areal interpolation to deal with differing regional structures in international research. *ISPRS International Journal of Geo-Information*, 9(2), 126. https://doi.org/10.3390/ijgi9020126.

Patuelli, R., Mussoni, M., Candela, G. (2016). The effects of world heritage sites on domestic Tourism: A spatial interaction model for Italy. In: Patuelli, R., Arbia, G. (eds) *Spatial Econometric Interaction Modelling. Advances in Spatial Science.* Springer, Cham. https://doi.org/10.1007/978-3-319-30196-9_13.

Şahin Körmeçli, P. (2024). Accessibility of urban tourism in historical areas: Analysis of UNESCO World Heritage Sites in Safranbolu. *Sustainability*, 16(6), 2485. https://doi.org/10.3390/su16062485.

UNESCO World Heritage Centre. (2024). *World Heritage List*. https://whc.unesco.org/en/list/.

doi.org/10.26398/asaproc.0075

"Inhumane": Quantitative textual analysis of a qualitative survey on gender violence among victims of international human trafficking

Micaela Arcaio, Federica Barbaro, Anna Maria Parroco

Department of Psychology, Educational Science and Human Movement, University of Palermo, Palermo, Italy

1. Introduction

Researchers often employ focus groups to gauge more information and insight on the phenomena they are studying (CDC, 2024). The qualitative interviews coming from this kind of research offer personal perceptions and experiences from stakeholders, thus providing breeding ground for novel interpretations of results that come from solely quantitative analyses. While usually either audiovisual records or their transcripts are employed to evaluate what is discussed during a focus group, in the last few years, several computational techniques, including machine learning and generative pre-trained transformers (GPTs), have helped develop quantitative analysis methods that process massive textual datasets (Nielbo et al, 2024). Text mining has been used to examine and find key themes in group interviews' transcripts to give objectivity to their analysis and natural language processing (NLP) techniques are used to first gather information on common themes, as well as to inform on the feelings risen from the text via more advanced sentiment analysis tools (i.e., Payton et al., 2018).

In this context, the aim of this work is to examine the transcript of a focus group about gender violence among victims of international human trafficking, using two text mining techniques. The focus group involved six key stakeholders working in the third sector with expertise on the matter – either by working with victims of gender violence, or by working with victims of human trafficking, or both – as well as a moderator. The stakeholders work in the province of Palermo (Italy) and refer to their experiences in this specific territory.

2. Human Trafficking and Gender-Based Violence

Human trafficking has been defined by the United Nations Office on Drugs and Crime in the Palermo Protocol as "the recruitment, transportation, transfer, harboring or receipt of people through force, fraud or deception, with the aim of exploiting them for profit" (UN, 2000). Violence and fraudulent promises of education and job opportunities are used to force or lure possible victims (UNODC, 2024).

No region in the world is exempt from this phenomenon, and people of all sexes and ages have been found to fall victims of this crime. The most significant transregional flow – i.e., from one region of the world to another, in contrast to domestic trafficking that happens within a certain region – concerns African people (UNODC, 2022). Moreover, 68% of trafficked individuals are moved by either governance-type crime groups or business-like crime groups: indeed, these kinds of organizations are able to traffic more people than individual traffickers can, as well as to retain them for longer periods of time, likely due to the extent of the violence used to coerce them (UNODC, 2022).

The detection of cases of human trafficking more often than not happens because of proactive

actions of the victims who manage to exit exploitation and thus self-report to the authorities: 41% of cases arise from victims' actions while only 9% emerge from other institutions such as hospitals, schools or NGOs (UNODC, 2022). Self-reporting can, however, be hindered by the perceptions of the victims themselves. Trafficked individuals struggle with identifying themselves as such due to psychological mechanisms such as denial or repression, inability to see themselves as victims, as well as lack of knowledge of their rights and the perception that their exploitation is happening because of their irregular migration status (Bloch and McKay, 2013; Van Meeteren and Hiah, 2020).

The exploitation of victims of human trafficking, thus, assumes several forms: in 2020, 38.8% of the detected cases were victims of forced labor and 38.7% were involved with sexual exploitation (UNODC, 2022). Women and girls constitute 60% of all detected victims. This statistic is followed by what can be found in case summaries: women are three times more likely to experience violence than men[1], which includes both physical and sexual violence (UNODC, 2022).

The venues in which this exploitation happens also matter. In the court case summaries for human trafficking in 2012-2022, 11% of victims were reported to work in the streets, while 18% were employed in apartments supplied by traffickers and 19% in brothels (UNODC, 2022).

Given this premise, the existence of a connection between human trafficking, gender-based violence (GBV) and sex work is clear. Sex work is defined as the supply of sexual encounters in exchange of money. Although many scholars debate on whether sex work is the result of women's exploitation or an act of sexual freedom, it has been recognized that there is a structural component to the exploitation and violence experienced by sex workers, with migrants, people living in poverty and women of color more likely to be abused than their white middle-to-upper class counterparts (Connelly, 2016). Even policies and laws decriminalizing prostitution still exclude migrants from their protection. For example, the Prostitution Reform Act issued in New Zealand in 2003 – that aimed to eliminate all laws against consensual sex work, as well as to improve workers' conditions by protecting them from exploitation and by promoting their well-being, health and safety – excluded migrants because it considered them as "seasonal workers" (Armstrong and Abel, 2020).

The three components of GBV – power, privilege and positionality[2] – are also shared with human trafficking. Power is exercised through violence by traffickers to force people into sex work; privilege and positionality determine the relationship between trafficker and trafficked (Ali and Rogers, 2023).

3. The focus group: the transcript and methods of analysis

The focus group whose transcript was analyzed in this study involved seven people: one moderator and six key stakeholders. The stakeholders included two experts in migration studies and agents in the third sector in the province of Palermo (Italy): two of them work within an association that provides psycho-social assistance to migrants to promote human rights and to fight GBV amongst victims of sex trafficking; another one works in an association whose aim is to contrast all GBV, and intimate partner violence in particular; the last participant in the focus group is a lawyer who works in an association that offers legal help to migrants.

The information sought during the interviews included: the victims' background, both in terms

[1] Court case summaries report that 31% of men have not experienced violence during their trafficking, whereas 79% of women are reported to have experienced extreme or explicit physical violence or threats and psychological violence (UNODC, 2022).

[2] In the literature on GBV power is defined as the ability to control whether an individual gets access to resources and opportunities as well as the ability to prevent another individual to make their own decisions. Privilege is due to the position of individuals in hypothetical hierarchies and it's strictly related to positionality, which refers to the placement within relationships themselves (Ali and Rogers, 2023).

socio-demographic characteristics and migration history; whether migrant victims of GBV had peculiar characteristics with respect to Italian victims; the profile of traffickers and perpetrators; what compels victims to report the abuse.

The record of the conversation was transcribed so that the textual analysis could be performed. It was kept in Italian to maintain the original language and vernacular used during the focus group. The text ended up being made of 341 entries (i.e., number of times the conversation shifted from one person to another) for a total of 13,266 words.

Prior to the analysis, the text was processed as following: numbers, common and uncommon stopwords, punctuation and extra spaces were removed; then it was transformed to find the stem of the words. After pre-processing, an exploratory wordcloud was created to check for the most common words in the text. Later a sentiment analysis was conducted. In this case, Saif Mohammad's NRC Emotion lexicon was used (Mohammad and Turney, 2013). This lexicon associates a list of words with eight basic emotions – anger, fear, anticipation, trust, surprise, sadness, joy, and disgust – and whether the overall sentiment is positive or negative. The choice for this particular lexicon was due to its availability in Italian.

The text was then evaluated qualitatively in conjunction with the quantitative results to gauge more information about human trafficking from the perspective of operators in the third sector.

The textual data used in this work was examined using R version 4.3.2 (R Core Team, 2023), and the packages tm version 0.7-13 (Feinerer and Hornik, 2024), wordcloud version 2.6 (Fellows, 2018), and syuzhet version 1.0.7 (Jockers, 2015).

4. Results

After the pre-processing of the transcript of the record of the focus group, a first exploratory wordcloud[3] was created (see Fig. 1). This graph offers a preview of the key concepts connected to the experiences lived by the victims of human trafficking. Here are some of the themes arisen:

- Violence and abuse: words such as *violenza*, *carnefici*, *disumane*, *vittima*, *agghiacciante* (i.e., violence, perpetrator, inhumane, victim, harrowing) highlight what victims of human trafficking live through;
- Women and minors: *donna*, *bambini*, *scuola* (i.e., woman, children, school) point to the peculiar vulnerability of these groups;
- Trafficking and exploitation: *attività*, *denunce*, *rete*, *problema*, *sessuale*, *clienti*, *strada*, *manager*, *sistemica*, *complesso* (i.e, activity, criminal report, network, problem, sexual, clients, street, manager, systemic, complex) all refer to the criminal nature of human trafficking and to the different forms of exploitation victims are subject to;
- Origin and destination: words like *ghanesi*, *cultura*, *mediare*, *italiano*, *nazionalità* (i.e., Ghanaian, culture, mediate, Italian, nationality) indicate how human trafficking is tied to migration flows and culture.

[3] The wordcloud in Fig. 1 is in Italian because it is the language of the text.

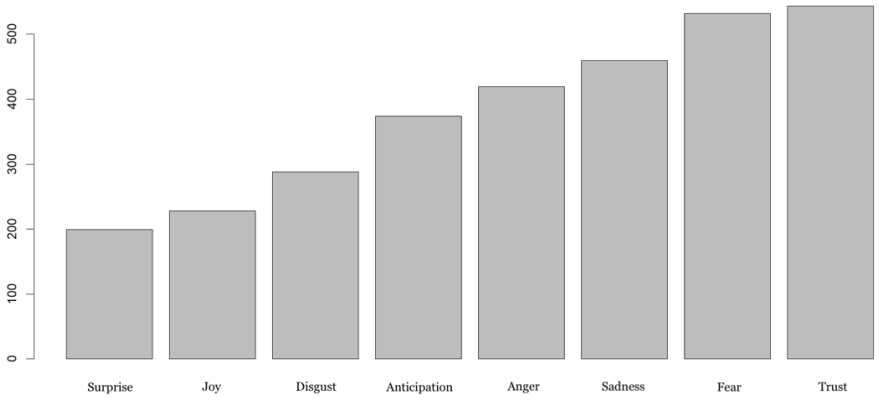
Fig 1 Wordcloud of the transcript of a focus group record



As for the results of the sentiment analysis (see Fig 2), they show a significant prevalence of negative emotions, such as sadness, fear, anger and disgust. This reflects the gravity of what victims of human trafficking experience, often including abuse, exploitation and physical and psychological violence.

The fact that trust is the most common sentiment is imputable to two mechanisms, one opposite the other. On the one hand, trust refers to the deception and tricks used to lure victims into trafficking, which leads them to the loss of autonomy and agency; on the other hand, trust is the foundation of the rapport between operators and victims seeking a way out.

Fig 2 Sentiment Analysis Results



5. Discussion

The results of the quantitative analysis – i.e. the main themes arisen from the common words in the text, in conjunction with the leading questions asked to the members of the focus group – were then used as guidelines to support the qualitative text analysis. Thus, several pieces of information on human trafficking were gathered:

- *Victims' background*. Words in the text (translated): nationality, culture, culturally, educat*, school, minors, woman, Ghanaian, country
    - Socio-demographic characteristics: human trafficking victims are usually from poor and remote villages, mostly presenting a low educational level. Even when trying to get out, education and independence – essential to be free of the exploitation – are rarely sought after. General distrust between different communities – heightened when people of different genders are present – hinders the possibility for women to even learn the language spoken in

the place they live in.
    - Migration background: while most victims of human trafficking in the province of Palermo do have a migration history, significative differences are found between ethnicities. Ghanaians, contrary to their Nigerian counterparts, are rarely involved in illicit activities. Most activities involving solicitation are, however, managed by criminal organizations while the "workers" are from Nigeria, Côte d'Ivoire, and Sri Lanka.
- *Peculiar characteristics of GBV among migrant women*. Words in the text (translated): network, victim, trafficking, children, school, perpetrators, inhumane, systemic, perceptions.
  The victims of human trafficking discussed in the focus group share a history of migration, which further increases their vulnerability to exploitation by criminal organizations. These crime groups present a highly organized and hierarchical structure, particularly evident when it comes to the Nigerian mob, which presents a pyramid configuration from chiefs to sex workers, managed by group leaders that also control contact with clients. This organization is a parastatal system, which offers services, assistance and aid – mainly monetarily. The systemic violence as well as women's exploitation in the street become harrowing experiences that cause the loss of identity in victims.
- *Asking for help*. Words in the text (translated): resignation, violence, normal, trust, police report, mediation, complex, relationships, sexual.
  It is usually a mixture of factors that pushes women into asking for help, such as experience of abuse and violence, precarious health conditions, economic and social difficulties, a need for legal support and social protection. "Normal" is quite often used in the text: it sometimes happens that because of cultural reasons, and because of the victims' lower educational level, they lack awareness about their rights and cognizance about the nature of what they are going through. The very first moment they try to ask for help, the aid relationship might fail: on the one hand, these people come from a complex system, made of abuse and exploitation, so they do not report traffickers (mostly because of fear); on the other hand, because they are already asking for help, building trust with the operators is crucial. But building trust could be hindered both by linguistic and cultural barriers. While the therapeutic value of narration is often employed to assist in case of intimate partner violence, this tool cannot always be used for victims of human trafficking because of the complexity and gravity of the abuse. However, long-term aid measures that focus on the autonomy of people trying to get out, such as job orientation (that also offers psychological and economic independence), can counter the forced subordination to traffickers.

6. Conclusions

The aim of this work was to use quantitative text analysis tools to evaluate the record's transcript of a focus group on GBV among victims of sex trafficking. This analysis highlighted how these victims mainly come from low socio-economic environments, heavily characterized by poverty, low education and multi-problematic families. In these contexts, migration is sought after for better living conditions and opportunities. Here it is easy that traffickers often affiliated with highly organized crime groups exploit these disadvantages in their favor.

The focus group showed how human trafficking is a systemic phenomenon, powered by less than effective migration laws and lack of care from statal institutions. Indeed, even though the conversation was centered around a criminal act, the lack of mentions to "police", "institutions" and "justice" – even though a lawyer was part of the group – really stands out. The only time institutions were mentioned at all was to highlight how very little interested in the phenomenon institutions are.

Using advanced methodologies such as text mining lends a hand into understanding and outlining underlying dynamics to complex phenomena discussed in focus groups. Visualizing texts through

wordclouds and sentiment analysis tools allows for a clearer interpretation of results.

This work, of course, does not lack limitations. The analysis is mainly exploratory and does not deal with more complex text analysis techniques that also study the relationship between words or methods that show how the tone shifts during the whole conversation. These methods could, however, represent a possibility of further development. Moreover, the focus group could be replicated, thus involving operators working in other provinces and members of governmental institutions to also gather their insights on the dynamics of human trafficking.

Zealand

References

Ali, P., Rogers, M. M. (2023). *Gender-Based Violence: A Comprehensive Guide*. Springer.

Armstrong, L., Abel, G. (2020). *Sex Work and the NewZealand Model: Decriminalisation and Social Change*. Policy Press.

Bloch, A., McKay, S. (2013). Hidden dishes—How food gets on to our plates: Undocumented migrants and the restaurant and takeaway sector. *Journal of Workplace Rights*, **17**(1).

Centers for Disease Control and Prevention (2018). *Data Collection Methods for Evaluation: Focus Groups.* cdc.gov/healthyyouth/evaluation/pdf/brief13.pdf

Connelly, L. (2016) Debates on prostitution: An introduction to feminist politics and their influence upon international policy and practice. In: Jacob, F., *Prostitution: A Companion of Mankind*. Peter Lang

Feinerer, I., Hornik, K. (2024). *tm: Text Mining Package*. R package version 0.7-13, CRAN.R-project.org/package=tm

Fellows, I. (2018). *wordcloud: Word Clouds*. R package version 2.6, CRAN.R-project.org/package=wordcloud

Jockers, M.L. (2015). *Syuzhet: Extract Sentiment and Plot Arcs from Text*. github.com/mjockers/syuzhet

Mohammad, S. M., Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, **29**(3), pp. 436-465.

Nielbo, K. L., Karsdorp, F., Wevers, M., Lassche, A., Baglini, R. B., Kestemont, M., Tahmasebi, N. (2024). Quantitative text analysis. *Nature Reviews Methods Primers*, **4**(1)

Payton, F. C., Yarger, L. K., & Pinter, A. T. (2018). Text mining mental health reports for issues impacting today's college students: qualitative study. *JMIR Mental Health*, **5**(4)

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R-project.org

United Nations Convention against Transnational Organized Crime, November 15, 2000, unodc.org/documents/treaties/UNTOC/Publications/TOC%20Convention/TOCebook-e.pdf

United Nations Office on Drugs and Crime (2022). *Global Report on Trafficking in Persons.* unodc.org/documents/data-and-analysis/glotip/2022/GLOTiP_2022_web.pdf

United Nations Office on Drugs and Crime (2024). *Human Trafficking*. unodc.org/unodc/en/human-trafficking/human-trafficking.html

Van Meeteren, M., Hiah, J. (2020). Self-identification of victimization of labor trafficking. *The Palgrave International Handbook of Human Trafficking*, pp. 1605-1618.

doi.org/10.26398/asaproc.0076

Keyword-enhanced summarization from scientific literature through an integrated approach

Massimo Aria[a,f], Corrado Cuccurullo[b,f], Luca D'Aniello[c,f],
Michelangelo Misuraca[e,f] , Maria Spano[a,f]

[a] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy.
[b]Department of Economics, University of Campania Luigi Vanvitelli, Capua, Italy.
[c]Department of Social Sciences, University of Naples Federico II, Naples, Italy
[e]Department of Business Administration and Law, University of Calabria, Cosenza, Italy
[f]K-Synth spin-off, University of Naples Federico II, Naples, Italy

1. Introduction

The exponential growth of textual data and publications has intensified the challenge for researchers to stay updated and identify impactful findings (Landdhuis, 2016). This underscores the need for tools that enhance information retrieval efficiency, enabling academics to quickly discern key contributions, limitations, and practical implications within scientific literature. Automatic Text Summarization (ATS) addresses this need by generating concise summaries that distill essential information, facilitating rapid comprehension without compromising the original content's integrity (Murad and Martin, 2007; Moratanch and Chitrakala, 2017). This paper explores ATS methodologies and introduces *Integrated Text Summarization* (ITS) (D'Aniello *et al.*, 2024)., a novel extractive algorithm tailored to scientific texts. By leveraging features such as keywords and title terms, ITS identifies and extracts sentences conveying critical insights. Comparative evaluations against TextRank and ChatGPT demonstrate ITS's accuracy and efficacy across diverse scientific articles.

2. Theoretical framework and related work

ATS methods are mainly classified into extractive and abstractive approaches. Extractive summarization selects and preserves original sentences based on their relevance, whereas abstractive summarization rephrases the content to create new sentences that convey the core ideas, mimicking human summarization (Nenkova and McKeown, 2011). Extractive methods offer distinct advantages over abstractive ones: (1) they ensure higher factual accuracy by directly utilizing sentences from the original text, minimizing errors and inconsistencies; (2) they are more computationally efficient, making them ideal for real-time applications and large-scale datasets (Gambhir and Gupta, 2017); (3) they require less training data and are simpler to implement, broadening their accessibility for diverse applications and research; (4) they better preserve the author's intent and style by retaining the original phrasing, which is particularly valuable in scientific and technical contexts where precision and fidelity are essential (Mani, 2001).

The application of ATS in scientific literature has been widely explored (Zaheer *et al.*, 2020; Koh *et al.*, 2022). While prior studies have primarily focused on text length reduction, summarizing entire scientific texts remains challenging due to high computational demands, even with advanced models like BART (Lewis *et al.*, 2019) and PEGASUS (Zhang *et al.*, 2020), which can only process a limited number of terms at a time. Deep learning advancements, particularly Transformers like BERT and GPT, have significantly enhanced text summarization by demonstrating exceptional capabilities in understanding and generating human-like text (Brown *et al.*, 2020). However, summarizing long scientific texts poses additional challenges, such as identifying key knowledge dispersed across various sections of an article. Effective summarization approaches must therefore account for the entire document's structure. While some methods (Chen and Bansal, 2018; Meng *et*

al., 2021) address these issues, they remain computationally intensive and often struggle to extract key knowledge accurately from the structured nature of scientific documents.

3. An integrated text summarization approach for scientific text synthesis

Given the limitations of abstractive summarization, this study focuses on extractive summarization methods, which are classified into three main approaches: (1) Statistical methods, selecting sentences based on properties like frequency; (2) Rule-based methods, relying on predefined rules for sentence selection; and (3) Fuzzy logic-based methods, using fuzzy logic principles for extraction. Among statistical approaches, network graph-based methods are particularly prominent. In these methods, sentences are represented as nodes, and edges connect sentences with a similarity above a specified threshold, measured by word overlap or semantic similarity (e.g., word embeddings). Centrality is one of the important measures associated with a node in a graph. The sentences with higher centrality scores are then included in the summary. Two widely used algorithms in this category are TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004). TextRank measures sentence similarity based on normalized word overlap and uses the PageRank algorithm for ranking, while LexRank employs a modified IDF cosine similarity function. The sentences can be ranked according to their degree of centrality or by using PageRank. The pairwise sentence similarity is measured for the entire set of documents passed to the function. Although both algorithms are effective for summarizing scientific texts (Aria et al., 2022), LexRank's pairwise similarity calculations can be computationally intensive for long documents. Consequently, TextRank was selected as the foundation for our proposed approach, balancing performance and computational efficiency.

Compared to existing extractive summarization methods, ITS is an innovative unsupervised approach tailored to the unique characteristics of scientific documents: their structured format, the authors' keywords, and key terms detected throughout the text using computational algorithms.Scientific documents are typically organized into sections or paragraphs, often adhering to the IMRaD structure: Introduction, outlining the context and proposal; Methods, detailing the analytical approaches; Results, presenting findings; and Discussion, interpreting the results and addressing their significance, limitations, and implications for future research. While not all authors strictly follow this structure, most articles can be categorized into these sections upon analysis. To maximize the relevance of extracted sentences, ITS applies its algorithm to each section, considering both the structural format and the full text. This approach enables the extraction of key sentences from each section, ensuring a comprehensive and context-aware summary of the document.

Authors carefully select keywords to encapsulate the core content of their scientific documents. This selection is critical, as keywords not only summarize the proposed knowledge but also influence the document's discoverability and reader engagement. Keywords significantly impact decisions to read or purchase a document and play a strategic role in its visibility in bibliographic databases like SCOPUS, where they help match search queries with relevant content. By summarizing the essence of their work within these keywords, authors enhance the likelihood that key sentences in the text align with these terms. To consider this strategic element, ITS integrates authors' keywords into the sentence extraction process. Additionally, ITS expands the authors' keyword list by employing two keyword extraction algorithms: Rapid Automatic Keyword Extraction (RAKE) (Rose *et al.*, 2010) and TextRank for Keyword Extraction (Mihalcea and Tarau, 2004). This dual approach ensures the inclusion of both explicitly provided and algorithmically identified key terms, enriching the summarization process with highly relevant content.

The process of identifying authors' keywords and terms extracted by the two keyword extraction algorithms was conducted iteratively for each document section. Keywords were then counted within each sentence to determine their significance during ITS scoring. Sentences with a higher number of keywords were given greater weight in the analysis by replicating them proportionally. For instance, a sentence containing three keywords would be replicated three times,

amplifying its influence in the final summary. Subsequently, the TextRank algorithm was applied, allowing ITS to include both authors' keywords and those automatically extracted.

The expanded keyword list enhances the identification of sentences likely to contain critical knowledge, increasing the effectiveness of the summarization process. To determine the number of sentences to include in the final summary, a flexible approach was adopted: the top 20% of sentences from each section, based on length, were selected.

All the analyses were performed in the R environment (v. 4.3.1). The *textrank* and *udpipe* packages were used for keywords-sentences identification and RAKE, respectively. *Ad hoc* algorithms were developed to iteratively retrieve keywords for each section and count their occurrences across sentences.

4. Data and experimental setup

The ITS was evaluated using ten distinct scientific articles (Table 1). Although the sample size was limited, this selection was guided by two key objectives. First, the articles spanned diverse subject areas, including healthcare and social sciences, enabling a robust assessment of ITS across various disciplines. Second, we contacted the first or corresponding authors of each article, requesting them to identify the most important sentences from their work. To ensure unbiased feedback, the authors were not informed about the purpose of the ITS and were simply asked to highlight at least one relevant sentence per section. This approach provided a *gold standard* for evaluating the sentences extracted by the ITS, allowing for a direct comparison of the algorithm's performance with expert judgment.

ID	Document	Authors' Keywords	# Sections	Sentences per section Mean ± SD
doc_01	D'Aniello *et al.* (2022)	Academic Health Centers; Health policy; Healthcare configurations; Scientific productivity; Research impact	11	18.8 ± 11.56
doc_02	Aria *et al.* (2023)	Tourism impact; Citizens' perceptions; Tourism development; Tourism sustainability; Structural equation models	9	26.8 ± 21.52
doc_03	Robinson *et al.* (2016)	Data Citation Index; Data sharing; Citation practices; Scholarly communication; Repositories	7	27 ± 13.33
doc_04	Robinson *et al.* (2014)	Altmetric.com; Twitter; Mendeley; altmetrics; social impact; coverage; Web 2.0	5	21 ± 11.47
doc_05	Aria *et al.* (2020)	Quality of life; Bibliometric analysis; Thematic analysis	7	53 ± 21.14
doc_06	Aria *et al.* (2022)	Text analytics; Topic detection; Thematic mapping	7	26.43 ± 9.03
doc_07	Della Corte *et al.* (2018)	Destination governance; Distrust; Trust	7	46.7 ± 21.52
doc_08	D'Aniello *et al.* (2018)	Dogs; Human emotional smell; Interspecies emotional transfer; Emotional communication; Dog's heart rate; Dog–human bond	15	12.9 ± 13.68
doc_09	Ciavolino *et al.* (2022)	Partial least squares; Structural equation modelling; PLS-SEM; Bibliometrics citation analysis; Bibliometrix R package	13	23 ± 21.83
doc_10	Adamo *et al.* (2023)	Keratotic oral lichen planus; Depression; Anxiety; Mood disorder; Pain	9	13.8 ± 14.45

Table 1. Scientific articles for data analysis

We extracted and organized the texts into a data frame for each article. We excluded sections like the abstract, acknowledgment, supplementary data, and bibliography, focusing on the core

content until the conclusion/discussion. Additionally, we collected all the authors' keywords.
After data retrieval, each article's text was imported into R for processing and analysis. The text pre-processing phase involved the following steps: (1) *Tokenization:* sentences were divided into individual terms or tokens; (2) *Part-of-Speech (PoS) Tagging:* grammatical roles (e.g., noun, verb) were assigned to each token within a sentence.; (3) *Lemmatization*: tokens were reduced to their root forms by removing prefixes and suffixes, ensuring accurate interpretation of meaning and context. This step leveraged PoS tagging for precision. For benchmarking, we applied the classical TextRank algorithm as a baseline and included summaries generated by ChatGPT (GPT-4) in the evaluation study. ChatGPT, known for its advanced contextual understanding and ability to synthesize complex texts, was included due to its growing adoption in scientific research for tasks such as information retrieval, text summarization, and content refinement. This comparison allowed us to assess the performance of ITS relative to established methods and advanced AI models. To obtain a summary from ChatGPT, the following prompt is used:

"Extract the 20% most relevant sentences for understanding the content of the attached scientific article for each section and subsection of the document. Exclude sections such as the abstract, acknowledgment, supplementary data, and bibliography."

5. Results and conclusion

Table 2 presents a comparison of sentence matching accuracy between the three different methods —TextRank, ITS, and GPT-4o — when aligned with human annotations across ten articles.

ID	Sentences annotated by author N	Sentences identified by TextRank N (%)	Sentences identified by the ITS N (%)	Sentences identified by GPT-4o N (%)
doc_01	35	12 (34.3%)	19 (54.3%)	**22 (62.9%)**
doc_02	42	13 (31%)	**14 (33.3%)**	12 (28.6%)
doc_03	17	6 (35.3%)	**7 (41.2%)**	**7 (41.2%)**
doc_04	8	**3 (37.5%)**	**3 (37.5%)**	**3 (37.5%)**
doc_05	30	13 (43.3%)	**17 (56.7%)**	8 (26.7%)
doc_06	21	8 (38.1%)	**12 (57.1%)**	3 (14.3%)
doc_07	25	11 (44%)	**16 (64%)**	3 (12%)
doc_08	33	9 (27.3%)	**13 (39.4%)**	5 (15.2%)
doc_09	36	5 (13.9%)	**12 (33.3%)**	4 (11.1%)
doc_10	30	6 (20%)	**8 (26.7%)**	5 (16.7%)
Median [IQR]	30 [22-34.5]	8.5 [6-11.8]	**12.5 [9-15.5]**	5 [3.25-7.75]

Table 2. Comparison of sentence matching accuracy among TextRank, ITS, and GPT-4o

Friedman test	χ^2	**df**	**p-value**
	10.3	2	**0.006****
Durbin-Conover test: Pairwise comparison		**Statistics**	**p-value**
ITS - TextRank		3.10	**0.006****
TextRank - GPT-4o		1.14	0.268
ITS - GPT-4o		4.24	**<0.001****

Table 3. *Results of applying the Friedman test to the TextRank, ITS, and GPT-4o, followed by the Durbin-Conover test for pairwise comparison.*
*p-value: * Significant 0.01 < p-value ≤ 0.05; ** Significant p-value ≤ 0.01*

The Friedman non-parametric test was applied to evaluate whether differences between summarization methods were statistically significant, followed by the Durbin-Conover post-hoc test for pairwise comparisons to investigate these differences further. Results are shown in Table 3. As shown in Table 3, the results of the Friedman test ($\chi^2 = 10.3$, df = 2, p-value = 0.006**) indicate a significant difference in sentence matching accuracy between the summarization methods. The pairwise comparison of methods further reveals that ITS significantly outperformed both TextRank (ITS vs TextRank: p-value = 0.006**) and GPT-4o (ITS vs GPT-4o: p-value < 0.001**). No significant difference, instead, was observed between TextRank and GPT-4o (TextRank vs. GPT-4o: p-value = 0.268). Findings provided that GPT-4o generally achieves more matched sentences with author annotations than TextRank, particularly in documents like doc_01(22-62.9%) and doc_03 (7-41.2%). However, its performance is inconsistent across the sample, with significant decreases in matching accuracy for several articles like doc_05 (8-26.7%), doc_06 (3-14.3%), and doc_07 (3-12%), where ITS—and sometimes even TextRank—outperforms it.

The ITS results highlighted the importance of integrating additional scientific key features to improve the reliability of knowledge extraction. Although some extracted sentences did not match the annotated ones, they offered important insights into the analyzed documents. The ITS approach can help quickly identify the core content of documents while also providing additional information that enriches the understanding of the scientific material. The development of sophisticated text summarization techniques holds great promise for scientific research by enabling the efficient synthesis of large volumes of literature, helping researchers stay up-to-date with the latest advancements in their fields.

Acknowledgment

The following research projects have financed this research: (1) PRIN 2022 SCIK-HEALTH (Project Code: 2022825Y5E - CUP: E53D23006110006); (2) PRIN 2022 PNRR The value of scientific production for patient care in Academic Health Science Centres (Project Code: P2022RF38Y - CUP: E53D23016650001).

References

Adamo, D., Calabria, E., ..., SIPMO (2023). Anxiety and depression in keratotic oral lichen planus: a multicentric study from the SIPMO. *Clinical Oral Investigations*, *27*(6), pp. 3057-3069.

Aria, M., Cuccurullo, C., D'Aniello, L., Misuraca, M., Spano, M. (2022). Text summarization of a scientific document: A comparison of extractive unsupervised methods. In *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data* (Vol. 1, pp. 67-73). VADISTAT Press/Edizioni Erranti.

Aria, M., Cuccurullo, C., D'Aniello, L., Misuraca, M., Spano M. (2022). Thematic analysis as a new culturomic tool: The social media coverage on COVID-19 pandemic in Italy. *Sustainability*, **14**(6), 3643.

Aria, M., D'Aniello, L., Della Corte, V., Pagliara, F. (2023). Balancing tourism and conservation: Analysing the sustainability of tourism in the city of Naples through citizen perspectives. *Quality & Quantity*, pp. 1-21.

Aria, M., Misuraca, M., Spano, M. (2020). Mapping the evolution of social research and data science on 30 years of social indicators research. *Social Indicators Research*, **149**, pp. 803-831.

Brown, T., Mann, B., Ryder, N., Subbiah, M., ..., Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, **33**, pp. 1877-1901.

Chen, Y. C., Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Ciavolino, E., Aria, M., Cheah, J. H., Roldán, J. L. (2022). A tale of PLS structural equation modelling: episode I-a bibliometrix citation analysis. *Social Indicators Research*, **164**(3), pp. 1323-1348.

D'Aniello, B., Semin, G. R., Alterisio, A., Aria, M., Scandurra, A. (2018). Interspecies transmission of emotional information via chemosignals: from humans to dogs (Canis lupus familiaris). *Animal Cognition*, **21**, pp. 67-78.

D'Aniello, L., Spano, M., Cuccurullo, C., Aria, M. (2022). Academic health centers' configurations, scientific productivity, and impact: Insights from the Italian setting. *Health Policy*, **126**(12), pp. 1317- 1323.

D'Aniello, L., Aria, M., Cuccurullo, C., Misuraca, M., Spano, M., (2024). Extracting knowledge from scientific literature with an integrated text summarization approach. In A. Dister, D. Longrée (eds.), *Mots competes textes déchiffrés (JADT24)* Presses Universitaires De Louvain Vol.1 (pp. 239-248) ISBN:978-2-39061-471-5

Della Corte, V., Aria, M., Del Gaudio, G. (2018). Strategic governance in tourist destinations. *International Journal of Tourism Research*, **20**(4), pp. 411-423.

Erkan, G., Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, **22**, pp. 457-479.

Gambhir, M., Gupta, V. (2017). Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review*, **47**, pp. 1-66.

Koh, H. Y., Ju, J., Liu, M., Pan, S. (2022). An empirical survey on long document summarization: Datasets, models and metrics. In *ACM Computing Surveys*. ACM. https://doi.org/10.1145/3545176

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, **535**(7612), pp. 457-458.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing.

Meng, R., Thaker, K., Zhang, L., Dong, Y., Yuan, X., Wang, T., He, D. (2021). Bringing structure into summaries: a faceted summarization dataset for long scientific documents. *arXiv preprint arXiv:2106.00130*.

Mihalcea, R., Tarau, P. (2004) TextRank: Bringing order into texts. In *Proceedings of the **2004** Conference on Empirical Methods in Natural Language Processing*, pp. 404–411.

Moratanch, N., Chitrakala, S. (2017). A survey on extractive text summarization. In *2017 International Conference on Computer, Communication and Signal Processing* (pp. 1-6). IEEE.

Murad, M. A. A., Martin, T. (2007). Similarity-based estimation for document summarization using fuzzy sets. *International Journal of Computer Science and Security*, **1**(4), pp. 1-12.

Nenkova, A., McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, **5**(2–3), pp. 103-233.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., Costas, R. (2014). New data, new possibilities: Exploring the insides of Altmetric.com. *El Profesional de la Información*, **23**(4), 359-366. https://doi.org/10.3145/epi.2014.jul.03

Robinson-García, N., Jiménez-Contreras, E., Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, **67**(12), pp. 2964-2975.

Rose, S., Engel, D., Cramer, N., Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, pp. 1-20.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ..., Ahmed, A. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Pprocessing Systems*, *33*, pp. 17283-17297.

Zhang, J., Zhao, Y., Saleh, M., Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR.

doi.org/10.26398/asaproc.0077

A composite indicator to measure attitude towards sustainable tourism by European citizens

Francesca Bassi , Manuela Scioni
Department of Statistical Sciences, University of Padua, Italy

1. Introduction

In this paper we propose a composite indicator (CI) to measure attitude towards sustainable tourism by European citizens using data collected with the Flash Eurobarometer survey 499, involving more than 25,000 European individuals living in the 27 European Union countries. The survey collected information on travelling behaviour, booking channels and information sources for travel preparation, reasons for selecting destinations, options and information on sustainable tourism, intentions to put in practice actions to preserve resources and the environment when on holidays.

In the reference literature, as well as in official statistics, it is possible to find many proposals of composite indicators to measure sustainability and circular economy even in the tourist sector, however, they are all based on data referred to territorial level. In this paper, we use data collected at individual level, on a representative sample of European citizens. Following the OECD standard methodology (OECD/European Union/EC-JRC, 2008), we propose a composite indicator to measure attitude towards sustainable tourism. The composite indicator is calculated at individual level and is then aggregated to perform further analyses: comparing European countries, identifying segments of tourists within and between European countries, explaining tourists' behaviour through appropriate analyses.

Many contributions may be found in the reference literature on the construction of indicators to measure sustainability of touristic destinations (Alfaro Navarro et al., 2019). Such proposals differ both in the objective and subject matter of the analysis and in the data sources and territories involved. Literature in this context focused mainly on tourist destinations, eventually involving tourists considering their level of satisfaction. Recent contributions using individual data to measure attitudes towards sustainable tourism referred to the SUS-STAT measurement scale proposed by Choi and Sirakaya-Turk (2005) or its modified versions, and applied to specific populations, either for socio-demographic characteristics and/or location. For example, Obradovic and Stojanovic (2021) applied the SUS-STAT scale to measure residents' attitude towards sustainable tourism development in the area of Gradac River in Serbia; Saparniene et al. (2022) used the same scale to measure attitude towards sustainable tourism in a youth population.

In this paper, we aim to build a composite indicator, a synthetic measure of attitude towards sustainable tourism, allowing comparisons in time and space, by considering intentions to sustainable behaviour while travelling or on holiday. To this scope, we use data collected in a representative sample of the 27 European Member States. Actions to protect the environment are of course more important than mere attitudes, however, actions depend on attitudes. Therefore, we consider that this paper, measuring attitudes, can give an important contribution to the scientific debates on sustainable tourism measurements and its determinants.

2. Data

The data analysed in this paper were collected by the survey Flash Eurobarometer 499, one of the many surveys conducted by the European Union institutions to monitor public opinion and other relevant issues from European citizens and enterprises. The survey was conducted in October 2021 and collected information on travel behaviour and the impact on it of the Covid-19 pandemic, booking channels and sources of information for travelling arrangements, reasons for selecting

destinations, options and information on sustainable tourism.

The sample consisted of 25,714 European citizens over 15, distributed in the 27 European Union (EU) Member States (MSs). Also information on socio-demographic characteristics of the respondents were collected (European Union, 2021). In this paper we consider the answers to nine questions referring to the willingness to change habits to be more sustainable while travelling or on holiday, these questions refer to taking over nine specific actions plus a question asking a general disposition to change behaviour while on vacation (Table 1)[1].

Table 1. Percentages of 'yes' responses to the ten binary questions in the sample of European citizens

	% of 'yes' responses
Q1 Consume locally sourced products on holiday	55.3
Q2 Reduce waste while on holiday	48.4
Q3 Take holidays outside the high tourist season	42.4
Q4 Travel to less visited destinations	40.0
Q5 Choose transport options based on ecological impact	35.5
Q6 Pay more to protect the natural environment	35.0
Q7 Reduce water usage on holiday	34.8
Q8 Contribute to carbon-offsetting activities	33.7
Q9 Pay more to the benefit of the local community	32.6
Q10 I am not prepared to change my habits	14.7

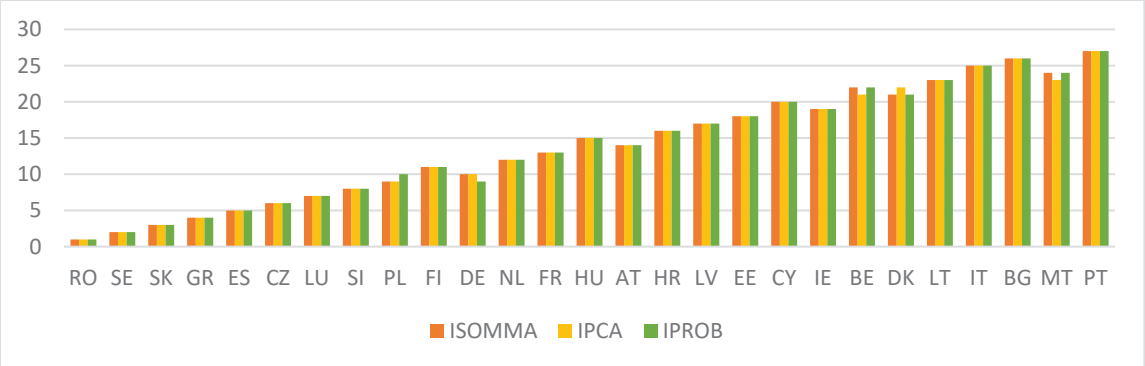
3. Indicator construction

Following the literature on constructing composite indicators (OECD/European Union/EC-JRC, 2008), we consider the first nine dichotomous variables reported in Table 1 to measure the attitude of European citizens towards sustainable tourism. The nine elementary indicators are aggregated through a weighted arithmetic mean.

Three approaches for weighting are considered: equal weights (ISUM), factorial weights obtained through a principal component analysis (1 component with eigenvalue > 1 explaining 46.8% of total variance, IPCA), and weights obtained as the inverse of the probabilities (IPROB) reported in Table 1. The idea behind this last method is that it gives more weight to the less common behaviours.

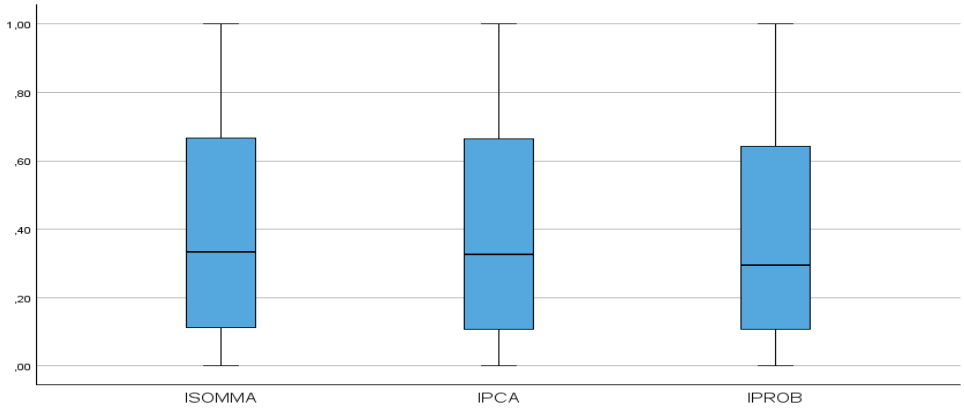
The three CIs give similar – but not exactly equal - results in terms of ranking of European countries (Figure 1), and are highly correlated and with similar values of their variation coefficient (see box plots in Figure 2), however, we selected the IPROB weighting scheme to conduct further analyses since we wanted to give more weight to those actions that are less likely to be performed and, as shown in Table 1, require to spend more money; we believe that declaring an intention to pay is correlated to a stronger attitude towards environment protection.

Figure 1. EU countries' ranking with the 3 CIs



[1] Descriptive analyses were conducted using post-stratification weights to ensure sample representativeness.

Figure 2. Boxplot of the three composite indicators

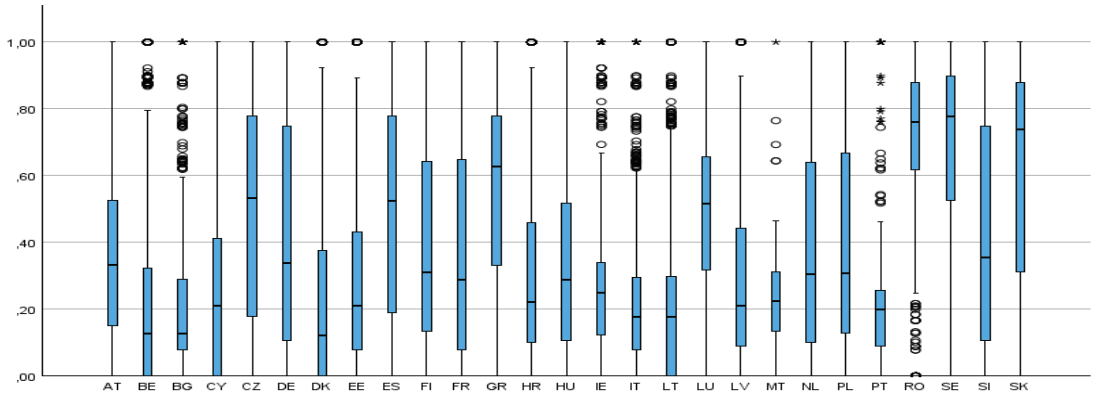


4 Statistical analyses

Having computed the individual composite indicator IPROB for each sample unit, many statistical analyses can be performed in order to explore determinants of attitude towards sustainable tourism. As, for example, Figure 3 shows a non-negligible variability both between and within European countries.

First of all, comparisons can be made between and within countries, i.e., both at citizens' and countries' level. The sociodemographic characteristics of the respondents are significantly associated to the level of attitude towards sustainable tourism, since mean values in groups by gender, age, occupation, family dimension, frequency of travelling, are significantly different. Then, in order to identify the drivers of this construct, we estimated a multilevel latent class regression model (Magidson et al., 2020) for composite indicator IPROB as dependent variable and tourists' sociodemographic characteristics as covariates; preferences about type of holiday, modes of organization of travelling and information sources enter the model as clusters' descriptors. Table 2 lists the results of the best fitting model.

Figure 3. Boxplot of the CI IPROB by country



Equation (1) describes a multilevel latent class regression model, i.e, a latent class regression model with a random intercept and a covariate (Magidson and Vermunt, 2006).

$$E\left(y_{ij}|X=k,z_{ij}\right)=\alpha_i+\beta_k z_{ij}$$
$$\alpha_i=\alpha+\lambda F_i$$

where y_{ij} denotes a continuous variable observed on individual i belonging to group j; X a

categorical latent variable with *K* classes; z_{ij} a covariate; α_i a latent class specific intercept, allowed to vary across individuals, the specific parameterization used for the intercept captures respondents' differences in their value for the dependent variable. λ is a factor loading and F_i distributed ad a standardised Normal. This implies that $E(\alpha_i) = \alpha$ and $V(\alpha_i) = \lambda^2$.

Table 2. Latent class multilevel regression model with IPROB as dependent and sociodemographic covariates.

	Cluster 4	**Cluster 3**	**Cluster 2**	**Cluster 1**
Countries	**CZ, GR, RO, SE**	**AT, BE, CY, FR, HU, IE, LU, NL**	**BG, EE, ES, FI, LT, LV, MT, PL, PT, SI**	**DE, DK, HR, IT**
Mean IPROB	13.2851	7.2319	7.1638	6.5606
Intercept	-0.2266*	0.2244*	0.4529*	-0.4507*
Age	-0.0181*	-0.0199*	-0.0192*	-0.0169*
Gender				
Male	-0.0242	-0.8128*	-0.6134*	-0.4860
Female	1.7189*	-0.1087	0.5526*	0.4283
Travelling before Covid				
Several times a year	1.4474*	0.7848*	0.8121*	1.8466*
Once or twice a year	1.4007*	0.3220	0.4057	0.7110
Once every few years	1.8167*	0.0721	-0.4440	0.3898
Never	1.3692*	-0.6096	-0.6989	-1.3588*
Occupation				
Self-employed	1.7449*	-0.0885	0.7558*	0.1534
Office worker	1.7238*	0.4844*	0.8179*	1.0255*
Manual worker	0.9950*	-0.0812	0.1300	-0.9058*
Without a professional activity	1.3473*	0.6325*	0.5326*	0.3062
Family members	-0.0022	-0.0300*	0.0070	-0.0060
Var(u_j)	2.1829*			
ICC	0.0563			

Four clusters of European countries are identified with different average values of the CI measuring attitude towards sustainable tourism (Figure 3). The considered sociodemographic characteristics of the respondents enter into the four multilevel regression models with different effects both in terms of significance, sign and magnitude. For example, while age of the respondent has a negative impact on the measure of attitude in all four clusters, never travelling before the pandemic has a positive effect in cluster 1 (the group of EU countries with the greatest attitude towards sustainable tourism) and a negative effect in cluster 4 (the group with the lowest attitude towards sustainable tourism). Family dimension has a statistically significant negative effect only in cluster 2. Table 3 reports clusters' profiles in terms of tourists' preferences. These figures profile the four clusters.

In EU countries of cluster 1, we find tourists who book mainly commercial accommodations thorough online platforms; they choose destination for the availability of activities and refer to many online sources for information about travelling, especially to reviews, ratings, but also personal experience and people recommendations; in this cluster, European citizens show the highest attitude towards sustainable tourism and appear also very active in searching for information and organizing holidays. In cluster 2, tourists choose mainly camping as type of accommodation and are characterised by attention to the price of the overall trip, they like blogs as a source of information. Tourists in cluster 3 are different form the others mainly because they use more traditional methods for booking and also traditional media for information (radio and tv). Finally, citizens in cluster 4 are those with the lowest score of the index, they look at cultural offerings and trust travel agencies.

Figure 4. Clusters of EU countries

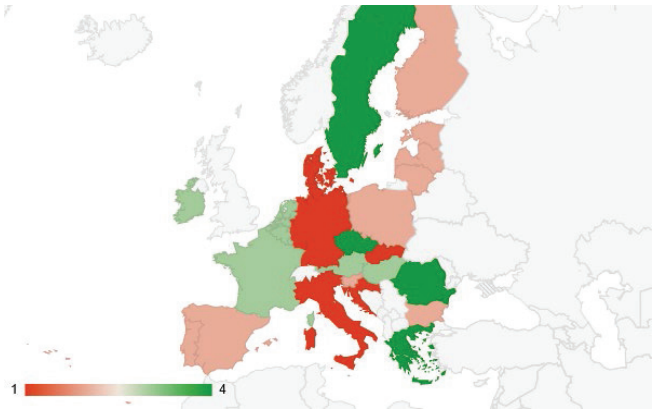


Table 3. Latent class multilevel regression model: cluster profiles %

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Accommodation					
Commercial accommodation with board included	55.7	39.7	46.1	42.2	45.6
Commercial accommodation without board	28.2	25.0	21.1	24.8	24.1
Private accommodation	39.7	32.2	34.1	31.4	34.2
Your own property or second home	25.9	17.0	14.6	16.3	17.7
Staying at friends' or relatives' place	35.9	26.7	26.0	23.9	27.8
A place organized via online platform no payment	7.9	7.9	4.8	8.1	6.8
Camp site or holiday village	16.9	17.6	14.1	13.7	15.6
Method to organize travel					
Online private house offering	24.6	18.6	21.1	19.0	20.8
Online platforms for professional accommodation	41.5	34.4	33.2	31.0	34.8
Online platforms combining travel services	34.1	27.1	28.0	26.4	28.7
Website of hotel, airline	26.4	28.3	19.9	18.7	23.3
Phone	18.0	14.3	13.0	15.6	14.9
Counter	15.5	16.0	17.0	17.5	16.5
Someone you know	29.7	21.4	19.3	21.2	22.2
At destination	15.4	12.9	9.7	11.4	12.0
Reason for choosing destination					
Cultural offerings	17.6	14.0	18.1	20.3	17.2
Natural environment	18.7	19.0	19.2	19.1	19.0
Activities available at destination	11.6	9.5	11.3	9.7	10.6
Price of overall trip	16.2	19.9	17.9	16.0	17.8
Accessibility for all	12.5	8.0	10.9	7.6	9.9
Promotion of eco-friendly practices	3.2	6.3	3.6	3.5	4.3
Destination reached by low-impact transport	4.1	3.7	1.9	3.4	3.1
Sustainability certification	1.9	3.1	2.2	2.5	2.4
Local population involved in tourism	2.8	4.0	2.1	3.4	3.0
Health and safety guidelines	7.9	7.6	7.6	6.6	7.5
Information sources					
Personal experience	44.4	37.1	33.8	38.4	37.5
Recommendations from people	71.3	46.5	63.0	53.6	58.6
Guidebooks or magazines	9.6	10.7	5.6	7.2	8.0
Website or social media	25.7	24.8	23.6	20.5	23.9
Travellers' reviews and ratings	50.8	31.5	42.3	33.2	39.6
Influencers and blogs	9.1	11.5	10.0	7.2	9.8
Promotion campaigns	5.0	4.7	5.0	4.3	4.8
Travel agencies	11.9	13.6	13.0	15.2	13.3
Tourism offices	9.4	12.6	9.1	10.0	10.3
Newspapers, radio, tv	10.6	11.2	6.7	10.6	9.3

5. Conclusions

In the realm of tourism studies, environmental sustainability has predominantly been assessed at the destination level, with composite indicators evaluating the impacts and performance of tourist locations and regions, whereas approaches that directly measure the environmental attitudes and behavioural intentions of individual travellers have received comparatively less attention. While assessing sustainability from the viewpoint of tourists is equally crucial, to date there exists a paucity of data explicitly gathered with the aim of quantifying tourists' environmental attitudes, awareness, and propensity for sustainable behaviour during their travels.

In this paper we propose a composite indicator (CI) to measure attitude towards sustainable tourism by European citizens using data collected with the Flash Eurobarometer survey 499. Data collected on attitude towards sustainable tourism is quite poor (10 binary variables), leaving little space for development in the construction of the composite indicator. However, some further research may be dedicated to identifying the best CI, especially testing alternative methods of weighting and aggregation. These methods should ground in the existing literature on sustainable touristic behaviour, especially exploiting the hypothesis that the willingness to perform different types of actions might be linked to different levels of environmental attitude, as well as trying to consider the willingness to perform more than one action jointly. Actions could also be weighted depending on the specific dimension (e.g., social, economic, or environmental) of tourism sustainability on which they are expected having an impact.

For what concerns model specification, second-level covariates, i.e., variables measuring phenomena at country level that, according to the reference literature may affect attitude towards sustainable tourism, may be added to the latent class multilevel model to evaluate improvement in fit and statistically significant effects. Random effects from the estimated model show non-negligible heterogeneity between countries.

Our analyses show that attitudes towards sustainable tourism are affected by socio-demographic characteristics as well as by preferences for accommodation types, reasons for destination choice, booking methods, and information sources used. It is very important to explore in detail the determinants of attitude towards sustainable tourism in order to device effective policies to support it; a good attitude will then translate in concrete actions to preserve the natural environment. An important result to be considered in design of policies is that drivers of attitude may different across segment of tourists and European countries, policies should be specifically tailored in this respect.

References

Alfaro Navarro, J.L., Andrés Martínez, M.E., Mondéjar Jiménez, J.A. (2019). An approach to measuring sustainable tourism at the local level in Europe. *Current Issues in Tourism*. 23: pp. 423–437.

Choi, H-S.C., Sirakaya-Turk, E. (2005). Measuring residents' attitude toward sustainable tourism: Development of sustainable tourism attitude scale. *Journal of Tourism Research*. 43: pp. 380–394.

European Union (2021). *Flash Eurobarometer 499. Report*.

Magidson, J., Vermunt, J.K., Madura, J. P. (2020). *Latent Class Analysis*, Sage.

Obradović, S., Stojanović, V. (2021). Measuring residents' attitude toward sustainable tourism development: a case study of the Gradac River gorge, Valjevo (Serbia). *Tourism Recreation Research*. 47: pp. 499–511.

OECD/European Union/EC-JRC (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing, Paris, https://doi.org/10.1787/9789264043466-en.

Šaparnienė, D., Mejerė, O., Raišutienė, J., Juknevičienė, V., Rupulevičienė, R. (2022). Expression of behavior and attitudes toward sustainable tourism in the youth population: A search for statistical types. *Sustainability*. 14: 473.

doi.org/10.26398/asaproc.0078

Data quality assessment framework for composite indicator construction: An application to service delivery

Francesca Bassi[b], Hillary Muhanguzi[a], , Manuela Scioni[b].
[a] Dept. of Statistical Methods and Actuarial Sciences, Makerere University, Kampala, Uganda.
[b] Department of Statistical Sciences, University of Padova, Padua, Italy.

1. Introduction

Establishing and sustaining elevated standards of service delivery, akin to its measurement is a vital priority for any nation. The initiation, packaging, solicitation, delivery and fulfilment of service needs involve diverse parties and interconnected iterative processes, rendering this phenomenon multi-dimensional. Efforts to assess service delivery using cross-sectional surveys and administrative data systems result in a diverse range of thematic indicators and statistics, which might occasionally foster non-convergence in policy discussions around service delivery.

Owing to this complexity, composite indicator (CI) methodologies utilising statistical aggregation and ranking provide effective means to capture and assess such a multi-faceted and data intensive phenomenon. Composite indexing entails mathematical aggregation of a collection of distinct indicators that measure multi-dimensional concepts but usually have no common units of measurement. Indeed, CIs have become increasingly significant in assessment of socio-economic phenomena related to service delivery, such as human development (United Nations Development Program [UNDP], 2024), access to services (Marozzi and Bolzan, 2018), Erasmus mobility effects (Fabbris and Scioni, 2021), etc. Despite the significant utility of CIs, their construction process is not methodologically straightforward (European Commission, Joint Research Centre and Organisation for Economic Co-operation and Development[OECD], 2008) as it prompts discourse concerning its design, the decisions necessitated at each phase, the individuals responsible for measurement, and importantly (though frequently neglected), the nature and quality of data integrated into model. Along these lines, Jimenez-Fernandez and Ruiz-Martos, (2020) assert that the methodological choices taken by any CI builder should be explicitly recognized. Albeit the methodological debates (Nardo et al., 2005), composite indicators remain vital due to their capacity to simplify complex measurement constructs and their political appeal. Amidst the criticisms, OECD outlines ten phases from theoretical framework to data visualization in their manual, which have served as a model for CI methodological paths for many years. The selection of data is a crucial stage, since its quality profoundly affects the reliability of the composite indicator.

Our primary objective is to construct a novel composite indicator of service delivery while acknowledging that the nature and quality of data used parallels the reliability of the CI. Moreover, we appreciate that numerous distinct elementary indicators assess service delivery, nevertheless, not all are suitable for inclusion in the CI model. The justification and qualification of elementary indicators necessitate a thoroughly documented criterion that not only addresses the concerns of data producers but should also be developed collaboratively with prospective users of the CI for the sake of transparency and parsimony.

We also observe that many researchers have documented criteria for selection of elementary indicators to build a CI, however, from the accessed literature, there is insufficient information on how the criteria were cultivated and applied in their respective case studies. This study strengthens and elaborates on the data selection, a fundamental phase in composite indexing upon which all succeeding stages rely. We establish explicit criteria for data suitability and encapsulate it within a data quality assessment framework (DQAF), a structure used to evaluate existing practices against

best practices (International Monetary Fund [IMF], 2003). Quality is the degree to which a collection of inherent characteristics of an object fulfils specified requirements (International Standardisation Organization [ISO], 2015). By extension, data quality fundamentally refers to the suitability of data for its designated purpose and user specifications (Uganda National Bureau of Standards [UNBS], 2012; Brackstone, 1999). Enhancing the transparency of the novel CI process, addressing user needs, and mitigating complaints necessitates a comprehensive explanation of input data quality via the DQAF.

We acknowledge that various comprehensive DQAFs, such as the 2018 United Nations Statistical Quality Assessment Framework, the IMF Data Quality Assessment Framework, the 2017 European Statistics Code of Practice, and the Quality Assurance Framework of the European Statistical System, exist in three or four tiers. However, these are overarching frameworks that may not fully bring out the quality issues for this specific case study. Instead, we settle with Huppes et al. (2014) that our formulated DQAF will encompass data quality needs for the continuous improvement of the CI of service delivery.

In the rest of the paper, we present the methodology employed to formulate the DQAF and implementation in the selection of elementary indicators (Section 2). In Section 3 we present the DQAF, while the last section contains the concluding remarks and prospects for future research.

2. Methodology for the formulation of the data quality assessment framework

Initially, the DQAF is formulated collaboratively with nineteen experts in data quality and potential users of the CI. This aligns with the claims of Strong et al. (1997), who argued that quality cannot be assessed by producers exclusive of consumers who choose and utilize products. During the preliminary stage of its definition, the literature was examined to identify twenty-three prospective data quality components (DQCs) for the assessment of data quality. DQCs are delineated in consistence with Veiga et al. (2017), as relevant and measurable attributes used to characterize the quality (or deficiency thereof) of data for a specific case study. We established that there was significant overlap in most of these DQCs that emerged from a literature search. Therefore, the experts drawn from the Ugandan National Statistics Office, government departments and universities were enlisted to select the most pertinent DQCs for the case study. These experts were administered an instrument containing the request to rate each of the 23 DQCs using a Likert scale (with scores 1-indicator should not be assessed against this component to 5-indicator must be assessed to this component). To customise the most suitable DQCs for this case study, frequencies were analysed to identify the top five preferred components. The second step involved the formulation of the data quality assessment framework. So, for each of the five DQCs identified in the previous phase, case study (service delivery CI) data quality requirements were detailed as shown in Table 1.

The selection of merely five components corresponds with authors such Chen and Lee, (2021), who utilized five DQCs in their study on selection of good quality data. Indeed, these five components, along with the criteria, were consolidated into a matrix of statistical information as illustrated in Weidema (1998). Given that certain data points or elementary indicators may not fulfil all quality standards, it was essential to establish data quality levels with distinctly defined bounds. The bounds were constructed to characterize the quality of each elementary indicator/data point with respect to the case study. The criteria were categorised into five quality levels each assigned any of the five numerical scores (1–5) that represent quality levels ranging from very good to very poor, in accordance with Chen and Lee (2021); Weidema (1998); and May and Brennan (2003). A 5X5 matrix was given, comprising a cross combination of the five DQCs as rows and their related quality criteria in five aligned columns associated with the scores. This generated matrix formed the data quality assessment framework for this case study.

The third stage involved application of the formulated DQAF to identify and select quality elementary indicators. Since the goal is to construct a service delivery composite indicator under five dimensions of education, health, water, agriculture and roads, we collated and profiled 103

Table 1: Elaboration of the case study requirements by data quality components and criteria

Data quality component	Key requirement for the case study	Quality criteria
Relevance: data serves to address the purposes for which they are sought by users, or which they are presented as addressing.	It is envisaged that the CI will be produced on an annual basis, therefore availability of annual series is important. It is the desire that the CI will be replicated at the global level, therefore indicators responding to international service delivery standards and SDGs are a priority. The indicator is more useful if it responds to national development aspirations as well as the stages of the constructed theoretical framework for the CI.	The indicator responds to a global (SDG), national development framework or sectoral programme.
		The indicator responds to theoretical framework (service provider-user interaction) of the case study (CI for service delivery).
		Statistics are produced for the reference year and published for the year 2021.
		The frequency of production (data & indicator update) is annual.
		The indicator responds to a known global or national service delivery standard.
Interpretability: data are easily understood and properly used by users	This CI prefers the use of already published data by national authorities in the form that it is required. An indicator published in a graph or table with accompanying information on how it was produced is important for enhancing interpretation and use. Additionally, an indicator published in the form that it is required is of priority to this case study.	The indicator is published in the main national report using a table or graph or map.
		The indicator is published with some form of narrative text about the contents of the table.
		The indicator is published in the main national report in a table or graph with some form of explanation on compilation practices.
		The indicator is conceptually clear at face value.
		The indicator is not proxy (published in the format that it is demanded).
Accuracy: source data and statistical techniques are sound, and statistical outputs sufficiently portray reality.	It is the desire that this CI is replicated across many countries, therefore priority is accorded to data collected from and for the relevant country from tested traditional sources. Indicators that represent judgements (although they respond to the service delivery phenomena) are less prioritized over the indicators obtained using objective measurements. This reduces dynamism of the indicator figures.	Data for compilation of the indicator are obtained from a representative population.
		Data are obtained using verifiable measurements (objective measurements and observations as compared to opinions and guesses).
		Data are obtained from the country of reference.
		Data for the indicator are known, and their scope can be traced.
		The source of data is traditional statistical source-survey, census or administrative.
Methodological soundness: the methodological basis for the statistics follows internationally accepted standards, guidelines, or good practices	Data compiled and aggregated using known international standards are preferred. The credibility of the producer, the similarity of timeliness for the numerator and denominator is important. In addition, data compiled for statistical purposes is assumed to be more methodologically sound.	All data for computation of the indicator are compiled using either national or international standards
		Data and indicators are produced by credible national authorities.
		Numerator and denominator are obtained from a single source for the same reference period.
		The original intent for compiling data and producing the indicator is for statistical purposes.
		The methods of computation of the indicator are known or can be traced.
Statistical adequacy: data selected represent valid and reliable measures.	This CI is to be presented by statistical regions, therefore, for every indicator and source, adequate number of observations should be generated to allow for such disaggregation.	The source of data yielded adequate number of observations to guarantee the desired disaggregation for the case study (statistical region).
		The response rate for the study/module is reported and is above 70%.
		Source data are obtained from and for the entire country.
		For survey data, the standard errors and design effect for the data points are reported and tend to zero and 1 respectively.

prospective elementary indicators for inclusion into the model. Profiling involved documentation of meta data for every indicator/data point from the reports in which they were published. Utilising the DQAF, each fundamental indicator was assessed by three experts based on

user experience to generate data quality ratings upon which a simple average rating for each elementary indicator was calculated. A choice was made to identify the qualifying elementary indicators that achieved a target quality score < 2.0. As endorsed by Rousseaux et al. (2001) as referenced in May and Brennan (2003), the World Health Organisation (WHO) proposes the utilisation of the target quality score for evaluating health data quality (WHO, 2022). The performance of this DQAF in relation to this case study was evaluated by calculating the descriptive statistics for each DQC and the proportion of elementary indicators that met the threshold. Correlations of the ratings by DQCs were analysed to identify potential overlaps.

3. The data quality assessment framework

The foremost five favoured DQCs for the service delivery case study, indicated by the 19 experts, were interpretability (15), methodological soundness (15), relevance (14), accuracy (13), and statistical adequacy (13). These can be succinctly delineated as presented in Table 1. Contextualization divides the selected data quality components into two categories: endogenous and exogenous. The endogenous category includes three DQCs (accuracy, adequacy, and methodological soundness) that are concerned with the quality of data and its intrinsic worth, primarily associated with data generation process. The exogenous category consists of two DQCs: interpretability and relevance, which pertain to user's use of the data. The integration of endogenous and exogenous categorisation aims to provide a more precise assessment by considering both the data user and the producer. The metadata for each elementary indication served as the foundation for the three raters to allocate it a cell in the DQAF presented in Table 2, thus producing its quality score. Application of the threshold resulted in the selection of 51/103 (48.6% acceptability rate) elementary indicators for inclusion into the model. Additionally, a uniform and high correlation (ρ>0.70) among the three DQCs of accuracy, methodological soundness and adequacy was established. Coincidentally, these dimensions fall into the data producer category. This means that elementary indicators that were found accurate or adequate are also characterized by methodological soundness. Conversely, the two DQCs in the data user category, namely relevance and interpretability, show a high level of correlation (0.76). The correlation between adequacy and relevance (0.45) and interpretability (0.51) are the lowest. These weak correlations can be evaluated in two ways: first as a desirable attribute as detailed by Odeny et al. (2023), who posited that a combination of DQCs with low correlation scores should measure distinct dimensions of data quality. Second, the low correlation scores across the exogenous and endogenous categories underscore the disconnect between data generation and use. While a data producer may satisfy statistical standards for the development of an indicator, this does not invariably translate into great usability for the user, a reason that data producers should periodically conduct user-satisfaction studies.

Table 2: Data quality assessment framework for the service delivery composite indicator

	Criterion and Score				
	1	**2**	**3**	**4**	**5**
ce	Statistics/ the indicator responds to a documented global development framework, service delivery standard or theoretical framework for the use case. The timing and frequency of production is annual	Statistics/ the indicator partly responds to a documented national development framework or sectoral programme. The timing and frequency of production is annual	Statistics/ the Indicator is a proxy for measurement of a global, or national development framework or any service delivery standard or policy. The timing and frequency of production is not annual	Statistics/ the Indicator does not respond to any of the accessed global, or national development framework or any service delivery standard or policy. The timing and frequency of production is not annual	Statistics/ the Indicator are neither produced nor accessible
tability	Statistics/ the indicator is presented in a national report using tables/charts accompanied by narrations of its meta data. The indicator is conceptually	Statistics/ the indicator is presented in a national report using tables/charts without narrations of its meta data. The indicator is conceptually clear and	Statistics/ the indicator is not presented in a national report; however, source data can facilitate its compilation. The indicator is conceptually	Statistics/ the indicator is not presented in the national report and source data cannot facilitate its computation. The indicator is not	Statistics/ the indicator is not accessed, and there is no source data at the country level

DQC	Criterion and Score				
	1	**2**	**3**	**4**	**5**
	clear and presented in a format that it is demanded	presented in a format that it is demanded	clear	conceptually clear	that can facilitate its computation.
Accuracy	All source data are adequately representative of the population, obtained from verified experiments in the country under study.	Some parts of the source data do not adequately represent the population, are obtained from verified experiments in the country under study	Source data are representative of the population, are obtained from qualified estimates, or historical data or models	Source data are not obtained from the country under study and are based on expert judgements, guesses, or international sources.	Source data for the indicator and its scope cannot be traced.
Methodological Soundness	All source data are collected from a single source and time and the indicator compiled based on international or national standards, and guidelines. The data are compiled by credible national authorities with the original intent being statistical	Some parts of the source data are collected from different sources and times; however, the indicator is compiled based on international or national standards, and guidelines	Data are compiled based on compiler's context from non-authorized statistical sources, but the original intent is statistical	The original intent for compilation of data is not statistical and the data are compiled by non-authorized national sources using compiler's own method	Compilation practices and scope for the data cannot be traced
Statistical Adequacy	Source of data is obtained for the whole country, generates adequate sample size to allow for the desired levels of disaggregation. The response rates, standard errors and design effects are reported	Source data obtained for the whole country generates inadequate sample size allowing limited disaggregation. Response rates, standard errors and design effects are reported	Source data are not obtained from and for the entire country with size not allowing the desired disaggregation. Response rates, standard errors and design effects are not reported	Source data are obtained from different area, other than the country of study, however allowing some form of disaggregation	The scope of source data is unknown and does not allow for any form of disaggregation.

4. Conclusion

This study introduced a comprehensive DQAF to identify and assess elementary indicators for the construction of composite indicators. The developed data quality criterion and reference rules, simplified into a checklist, guides assessors in assigning objective quality scores to the elementary indicators and datasets, if accessible. The aggregated scores, along with the narrative rationale, assist users of CI to appreciate the methodological decisions underlying the selection of the elementary indicators, their quality and potential enhancements. The involvement of experts in selection of DQCs, and the actual scoring process may be more efficient than unilateral decisions, since it fosters ownership and reduces bias. A significant discovery is that the reference rules can consistently be modified to align with the specifications of a well-defined case study. This DQAF provided higher quality scores to elementary indicators that align with national and international development frameworks, such as the Sustainable Development Goals (SDGs), and those that yield annual data. Conversely, in its adaptation, other constructors who may favour elementary indicators that correspond to a sub-national program and receive monthly updates could need to allocate greater quality scores to these elementary indicators accordingly. The selection of assessors must acknowledge that data quality concerns require fundamental understanding of statistical procedures. This study reveals that certain DQCs are correlated, potentially leading to repetition and overlap in assessments. Therefore, DQCs exhibiting uniform or high correlations (for example accuracy and methodological soundness) may be amalgamated or alternative DQCs may be selected. Future research should focus on selecting alternative DQCs in these instances until an optimal uncorrelated combination is achieved to broadly reveal and characterize the data quality of elementary indicators.

References

Brackstone, G. (1999). DNSS Document: Managing data quality in a statistical agency. *Survey Methodology* , *25*(2), 139–149. https://unstats.un.org/unsd/dnss/print.aspx?docID=194

Chen, X., Lee, J. (2021). The identification and selection of good quality data using pedigree matrix. *Smart Innovation, Systems and Technologies*, *200*, pp. 13–25. https://doi.org/10.1007/978-981-15-8131-1_2

European Commission. Joint Research Centre., Organisation for Economic Co-operation and Development. (2008). *Handbook on Constructing Composite Indicators : Methodology and User Guide.* OECD.

Fabbris, L., Scioni, M. (2021). Pooling rankings to obtain a set of scores for a composite indicator of Erasmus + Mobility effects. *Social Indicators Research*, *156*(2–3), pp. 481–497. https://doi.org/10.1007/s11205-020-02513-6

Huppes, G., Nico W. van den, B., Erwin W., L., Bernhard L. van der, V., M. Nicoline, W. (2014). *Quality Assessment for LCA* (Issue January 1999). https://www.researchgate.net/publication/28649054_Quality_assessment_for_LCA

International Monetary Fund. (2006). *Data Quality Assessment Framework DQAF*. *August*, 2. http://dsbb.imf.org/images/pdfs/dqrs_factsheet.pdf

ISO. (2015). Quality management systems - Fundamentals and vocabulary. In *English: Vol. ISO 9000* (ISO 9000:2005 (E)). www.iso.org

Jimenez-Fernandez, E., Ruiz-Martos, M. J. (2020). Review of some statistical methods for constructing composite indicators. *Estudios de Economia Aplicada*, *38*(1), pp. 1–15. https://doi.org/10.25115/eea.v38i1.3002

Marozzi, M., Bolzan, M. (2018). An index of household accessibility to basic services: A study of Italian regions. *Social Indicators Research*, *136*(3), pp. 1237–1250. https://doi.org/10.1007/s11205-016-1440-0

May, J. R., Brennan, D. J. (2003). Application of data quality assessment methods to an LCA of electricity generation. *International Journal of Life Cycle Assessment*, *8*(4), pp. 215–225. https://doi.org/10.1007/BF02978474

Nardo, M., Saisana, M., Saltelli, A.(2005). Tools for composite indicators building. *European Commission Joint Reserach Centre*, *EUR 21682 EN*. http://farmweb.jrc.cec.eu.int/ci/bibliography.htm

Odeny, B. M., Njoroge, A., Gloyd, S., Hughes, J. P., Wagenaar, B. H., Odhiambo, J., Nyagah, L. M., Manya, A., Oghera, O. W., Puttkammer, N. (2023). Development of novel composite data quality scores to evaluate facility-level data quality in electronic data in Kenya: a nationwide retrospective cohort study. *BMC Health Services Research*, *23*(1), pp. 1–15. https://doi.org/10.1186/s12913-023-10133-2

Strong, D. M., Lee, Y. W., Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), pp. 103–110. https://doi.org/10.1145/253769.253804

Uganda National Bureau of Standards. (2012). *Uganda Standard: Code of Practice for official statistics* (First edit). www.unbs.go.ug

United Nations Development Programs. (2024). *Human Development Report 2023/24*. https://www.undp.org/sites/g/files/zskgke326/files/2024-03/hdr2023-24overviewen.pdf

Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLoS ONE*, *12*(6), pp. 1–20. https://doi.org/10.1371/journal.pone.0178731

WHO. (2022). *Data Quality Assurance. Module 1. Framework and Metrics*. https://www.who.int/data/ data-collection-tools/health-service-data/data-quality-assurance-dqa

doi.org/10.26398/asaproc.0079

Exploring art market trends: A Delphi study on collectible design

Laura Benedan[a], Federica Codignola[a], Carlotta Galeone[b], Paolo Mariani[a]
[a] Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Milan, Italy.
[b] Biostatistics & Outcome Research, Statinfo, Milan, Italy.

1. Introduction

Creative and cultural industries encompass firms that produce and exchange goods of elevated symbolic rather than material value, characterised by artistic creativity. These goods are also often characterised by uniqueness or limited reproducibility (Caves, 2000; Hirsch, 1972).

Considering anecdotal evidence based on publicly available data from the market and the media, the practice of collecting design is growing while becoming a considerable trend. Design objects belong to the aforementioned category of goods, and when they reach the typical criteria of collectibility (e.g., rarity, uniqueness, authenticity, etc.), they can be understood under a general conceptual category of *collectible design*. As these objects are traded through commercial platforms, this category can also be understood as a market category.

In particular, these objects circulate and are traded across and through typical art market platforms (auction houses, commercial galleries and their art fair stands, dealers, etc.), as well as across several commercial platforms outside the art system (flea markets, individual private exchanges, resale e-commerce websites, etc.).

Although there are no specific collectible design indexes, some art indexes include these items and show an increase in auction selling prices of collectible design items (e.g., Artprice, 2014-2024; Bonhams, 2014-2024; Cambi Aste, 2014-2024; Christie's, 2014-2024; Phillips, 2014-2024; Sotheby's, 2014-2024; etc.).

Another indicator of the rise in demand for collectible design items emerges from a general increase in the number of commercial or institutional actors (e.g., the public sector) that are specialised, based, or focused on collectible design (e.g. private galleries, auction houses and sales, fairs, e-commerce platforms, museums, special exhibitions, etc.). Moreover, an emerging body of anecdotal data from diverse media outlets indicates a rising interest in collectible design, as evidenced by an increase in references to this subject matter across a range of digital and traditional media, including social media, magazines, periodicals, and informal conversations.

Such established evidence of the growing importance of this category prompts further investigation in this trending field, especially given the lack of research on the category itself.

2. Collectible design

Wu (2013, p. 3) defined *20th-century design* as a "collection of functional and collectible objects dated between 1880 and 1980" and described *collectible objects* as assets "traded on an active and transparent resale market with a potential to realise financial gains or storage of value".

Bertacchini and Friel (2014) identified the segment of *collectible furniture design*, referencing both modern and contemporary design items that were originally conceived as functional goods and, in recent decades, have been integrated into the art domain through special collections and exhibitions in museums. The authors affirm that with the emergence of

demand by collectors, a specific market for collectible design has emerged. Hence, we recognise collectible design as a hybrid category.

While collectible design can be perceived as an independent market (Bertacchini and Friel, 2014), it also performs through the art market, in accordance with Robertson's (2005) art market categorization.

In this paper, we conceptualise collectible design as a *hybrid market category* encompassing specific *sectors* (e.g., furniture, glassware), *types* (e.g., Modern, Art Deco, Memphis), and *commodities* (e.g., the Wishbone Chair, designed by Hans Wegner in 1949 for Carl Hansen & Søn).

In this study, we also refer to collectible design as a category that includes both historical and contemporary design commodities produced in limited editions or as unique pieces, while excluding classic design (antique furniture and decorative commodities). Indeed, in accordance with the conventional art market classifications, classic design represents a distinct category, which is usually treated separately.

Although visual art consumption may be related to hedonic (Hirschman and Holbrook, 1982; Holbrook and Hirschman, 1982), aesthetic (Holbrook, 1980), and symbolic consumption (Hirschman, 1981; Hirschman, 1986; Park et al., 1986), design commodities are functional by nature. However, in this study, we argue that collectible design users (e.g., collectors), besides valuing these commodities for their functional and utilitarian benefits, may also be motivated by the need to satisfy their emotional, aesthetic, or symbolic needs connected, for instance, to self-image and social identification. In fact, collectible design mostly performs through the art market, which, owing to some of its peculiar characteristics, is perceived as extremely symbolic (Codignola, 2019; Graw, 2009; Velthuis, 2005, among others). This undoubtedly contributes to the symbolic significance of collectible design.

Consequently, by referring to collectible design as a hybrid market category of hybrid commodities existing in a hybrid space that blurs the lines between function and symbolism, this study seeks to understand whether the symbolic value of collectible design represents the main driver of attractiveness. If this is the case it would have a positive influence on economic performance and value (Codignola and Mariani, 2021; Throsby, 2008; Velthuis, 2005).

Upon this thesis is based the main objective of this study, which is to make a first attempt to define and contextualise collectible design.

3. Methodology

To make an initial attempt to define and contextualise collectible design, this study employs the Delphi method, a structured communication technique originally developed as a systematic, interactive forecasting tool that relies on a panel of experts (Dalkey and Helmer, 1963). The Delphi process involved multiple rounds of questionnaires, with feedback provided to participants between rounds to refine their responses and to move toward consensus.

To initiate the study, three in-depth qualitative interviews were conducted with experts in collectible design: two from Italy and one from France. These interviews aimed to gather in-depth insights into the current state and future direction of collectible design. Data collected from these interviews were subjected to both qualitative and quantitative analyses to extract meaningful patterns and trends.

The interviews were conducted in a structured format that allowed consistent responses across participants. Each interview was recorded to ensure its accuracy and completeness. After recording the interviews, the interviews were transcribed verbatim. As one of the interviews was conducted in French, it was translated into Italian before proceeding with the analysis, to maintain consistency in the language of analysis.

4. Data analysis process and preliminary findings

The process of analysis was structured as follows:

1) A data matrix was created from the transcribed responses of three experts. To focus on content-bearing terms, non-essential words, such as conjunctions, articles, and prepositions, were removed.

2) A lemmatisation process was applied, reducing words to their base forms or lemmas. For instance, verbs were transformed into their infinitive forms, and nouns were standardised to the masculine singular. This step was crucial for minimising redundancy and ensuring that words with the same root were treated as single entities in the analysis. The lemmatisation process resulted in a matrix comprising more than one hundred unique words (rows) across the three interviews (columns). This matrix served as the foundation for the subsequent quantitative analysis. A subset of the data matrix is illustrated in Table 1 as an illustrative example of a decreasing occurrence vocabulary.

Table 1. Example of a subset of the data matrix, frequencies, 2024

Words	Int. 1	Int. 2	Int. 3	Occurrences
original	8	3	8	19
person	5	6	7	18
interest	2	8	7	17
functional	9	0	7	16
furnishing	0	9	5	14
unique	9	2	2	13
culture	6	6	0	12
thinking	7	4	1	12
mind	2	9	1	12
value	1	10	1	12
…				

Source: data from Delphi study on collectible design

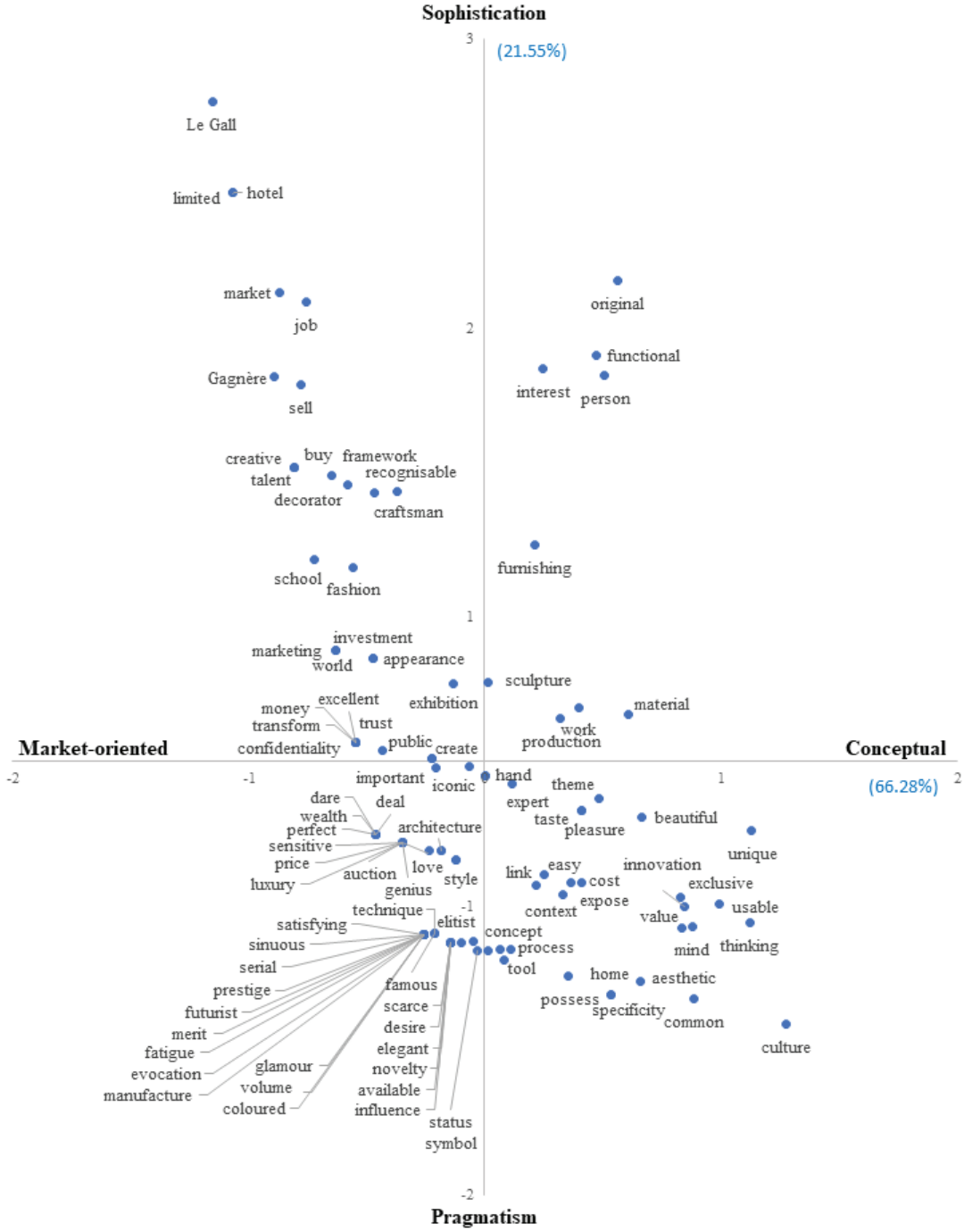
3) A principal component analysis (PCA) was conducted to identify the underlying patterns and relationships within the data, as well as to ascertain the convergence and divergence in perspectives on the topic of collectible design. This statistical method is particularly suited for the analysis of qualitative data, such as interviews, by condensing numerous variables—like words, phrases, and themes—into a smaller set of principal components that capture essential information. This technique identifies latent connections, reveals underlying structures, and enables visual representation of data, aiding in the interpretation of relationships and trends. Additionally, PCA helps focus on relevant factors while minimising noise, simplifying the analysis of complex qualitative datasets (Bolasco, 1999).

The PCA conducted on the words from the interviews revealed two primary dimensions that effectively organised the language used by the interviewees into meaningful themes. The results show that the first principal component accounts for a substantial proportion of the data's variability, explaining over 66.3% of the variance. This suggests that the first component is the most dominant factor in explaining the underlying patterns in the data.

The second principal component, which explains 21.6% of the variance, is also significant, though it represents a smaller portion in comparison to the first component. Collectively, the first two components explain 87.8% of the total variance, indicating that these two components provide a comprehensive representation of the dataset.

Figure 1 displays the first extraction of the Cartesian plane of concepts derived from the regression scores of individual words obtained through PCA of the three structured interviews. Each point represents a concept positioned according to the principal components that capture the greatest variance in responses. The plot highlights the relationships between the interview responses, showing how perspectives on collectible design converge or diverge.

Figure 1. Cartesian plane derived from PCA, regression scores, 2024



Source: elaboration from research data, Delphi study on collectible design

The left side of the horizontal axis is characterised by a focus on *market-oriented* terminology, including concepts such as “marketing”, “sell”, “buy”, “price” and “auction”. These concepts are closely associated with commercial activity, the value of transactions, and consumer behaviour, with a particular focus on tangible exchanges and measurable outcomes.

The *conceptual* terms on the right side of the spectrum, such as "unique", "thinking", "mind", and "value" prioritise abstract qualities like creativity, originality, intellectual depth, and intrinsic worth. The dichotomy reflects a tension between external, economic motivations and internal, ideational aspirations.

The vertical axis represents a continuum, with *sophistication* at the upper end and *pragmatism* at the lower end. The 'sophistication' pole is characterised by a set of attributes including "Le Gall", "limited", "original" and "recognisable". These traits emphasise exclusivity, refinement and intellectual or artistic elevation. In contrast, pragmatism is characterised by terms such as "available", "novelty", "common", "home", "manufacture" and "tool", which emphasise practical functionality, accessibility and widespread applicability. This vertical dimension reflects the contrast between elevated, niche approaches and grounded, utilitarian concerns.

5. Discussion and conclusion

The qualitative analysis of the expert interviews was instrumental in shaping and refining the Delphi questionnaire. By meticulously examining the insights and perspectives shared by these experts, we were able to identify the core themes that were central to our topic. This enabled us to design a questionnaire that effectively captured the most pertinent issues.

The early integration of qualitative methods into the Delphi process sets this approach apart. In particular, the combination of qualitative and quantitative methods, including lemmatisation and PCA, provides a nuanced understanding of the perspectives shared by experts.

This systematic approach ensured that the analysis was both rigorous and insightful, offering valuable contributions to better framing and understanding the collectible design category. Based on the insights gained from qualitative interviews, a comprehensive questionnaire was developed and administered to a panel of experts for the Delphi study.

The upcoming phases of the Delphi study will delve deeper into the factors influencing perceptions of collectible design. For instance, by refining the consensus among experts, this study offers a nuanced understanding of the key elements that allow a design item to be perceived as *collectible*.

In terms of implications for research, the findings will contribute to defining and contextualising more precisely, while facilitating a better understanding of the role of the symbolic value of these commodities in collecting patterns and users' perceptions.

With regard to the implications for stakeholders, the insights generated will serve as valuable resources for designers, collectors, and investors, enabling them to navigate and shape the evolving collectible design category. This knowledge will empower them to create and purchase items that not only resonate with current market trends but also retain their relevance and value over time.

References

Artprice. (2014-2024). www.artprice.com.

Bertacchini, E., Friel, M. (2014). Determining the value of modern and contemporary furniture design: An exploratory investigation, Department of Economics and Statistics Cognetti de Martiis. Working Papers 201437. University of Turin.

Bolasco, S. (1999). *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Carocci, Roma.

Bonhams. (2014-2024). www.bonhams.com.

Cambi Aste. (2014-2024). www.cambiaste.com.

Caves, R.E. (2000). *Creative Industries: Contracts between Art and Commerce*. Harvard University Press, Cambridge, (MA).

Christie's. (2014-2024). www.christies.com.
Codignola, F. (2019). The globalization of the art market: A cross-cultural perspective where local features meet global circuits, in M. Khosrow-Pour (ed). *Brand Culture and Identity: Concepts, Methodologies, Tools, and Applications*, Information Resources Management Association (USA), IGI Global, pp. 1327-1345.
Codignola, F., Mariani, P. (2022). Investigating preferences in art collecting: The case of the François Pinault Collection. *Italian Journal of Marketing*, pp. 107-133.
Dalkey, N., Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science,* **9**(3), pp. 458-467.
Graw, I. (2009). *High Price: Art between the Market and Celebrity Culture*. Sternberg Press.
Hirsch, P. (1972). Processing fads and fashions: An organization-set analysis of cultural industry systems. *The American Journal of Sociology*, **77**, pp. 639-659.
Hirschman, E.C. (1981). Comprehending symbolic consumption: Three theoretical issues, in E.C. Hirschman and M.B. Holbrook (eds). *Symbolic Consumer Behavior*, Association for Consumer Research, New York, (NY), pp. 4-6.
Hirschman, E.C. (1986). The creation of product symbolism. *Advances in Consumer Research*, **13**, pp. 327-331.
Hirschman, E.C., Holbrook, M.B. (1982). Hedonic consumption: Emerging concepts, methods and propositions. *Journal of Marketing*, **46**, pp. 92-101.
Holbrook, M.B. (1980). Some preliminary notes on research into consumer aesthetics, in J.C. Olson (ed). *Advances in Consumer Research*, Association for Consumer Research, Ann Arbor, (MI).
Holbrook, M., Hirschman, E.C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings and fun. *Journal of Consumer Research*, **9**, pp. 132-140.
Park, C.W., Jaworsky, B.J., MacInnis D.J. (1986). Strategic brand concept-image management. *Journal of Marketing*, **50**, pp. 135-145.
Phillips. (2014-2024). www.phillips.com.
Throsby, D. (2008). *The Economics of Cultural Policy*. Cambridge University Press.
Robertson, I. (2005). *Understanding International Art Markets and Management*. Routledge.
Sotheby's. (2014-2024). www.sotheby's.com.
Velthuis, O. (2005). *Talking Prices: Symbolic Meanings of Prices on the Market for Contemporary Art*. Princeton University Press, Princeton, (NJ).
Wu, L.T.J. (2013). *Financial Returns and Price Determinants in the 20th Century Design Market: Evidence from Sotheby's New York, 2007-2013 & 1966-2013*. Master's degree in art business Thesis, Sotheby's Institute of Art.

doi.org/10.26398/asaproc.0080

Resilience versus depression of Italian youth after the COVID-19 pandemic

Luigi Bollani[1], Clelia Cascella[2], Luigi Fabbris[3]
[1] Dept. of Economics, Social Studies, Applied Mathematics & Statistics, Univ. of Turin, Italy
[2] Invalsi, Rome, Italy
[3] Tolomeo Studi e Ricerche, Padua, Italy

1. Introduction

The theory of positive psychological capital ("PsyCap") includes resilience as one of the components of an individual's capital, defined as their ability to systematically develop competencies (Luthar and Cicchetti, 2000). The other components are optimism, hope, and self-efficacy. The complex structure of resilience makes it a relevant ability in case of social shocks, such as the COVID-19 pandemic. Conversely, depression, along with other psychopathologies (e.g. negative mood, confusion, severe anxiety, and suicide attempts) represent the antithesis of resilience. This suggests that resilient individuals are less susceptible to depression, while those who are depressed exhibit low or no resilience (Barzilay et al., 2020; Brooks et al., 2020). Preston et al. (2023), reviewing the literature on PsyCap, suggest that the COVID-19 pandemic exacerbated the mental health crisis among young people and may have raised their risk of suicide.

In order to explore the consequences of this health crisis and the potential for a swift return to normalcy, we conducted a survey of Italian adults in the second half of 2021 and 2022, when the COVID-19 pandemic was nearing its end. From the responses obtained, a sample of young adults (aged 18-34) was extracted. The dataset also included measures of resilience and depression, both assessed using psychological scales from the literature. In this paper, we analyse the relationship between these two constructs.

The research questions addressed in this study are:
Q1: To what extent is the level of resilience among Italians inversely correlated with their depression rate?
Q2: Are there side variables that can predict the relationship between resilience and depression?
Q3: Are there specific social groups that exhibit distinctive patterns of resilience and depression?

The rest of the paper is organised as follows: Section 2 introduces the data and the model for statistical analysis. Section 3 presents the main results of the statistical analysis of the data. Finally, Section 4 discusses the findings in relation to the existing literature and concludes the paper.

2. Data and methods

2.1 Data

A sample of Italian adults was surveyed between June and November 2021, and again in March 2022, using the computer-assisted web-based interviewing (CAWI) technique. A total of 1,192 responses were collected. After excluding cases with complete non-response or only one answer on the resilience question, the sample was reduced to 1,185 valid questionnaires. After excluding 7 respondents residing abroad and 1 individual who did not provide information on gender, we were left with 1,177 valid responses from Italian adults. Of these, 550 were young adults (aged 18-34), 518 were of working age (35-64), and 108 were aged 65 or older.

The survey was part of the RECALL project (*Research on the Effects of COVID-19 on All Lifestyles and social Links*), which aimed to study the wellbeing of Italian adults following the pandemic. The project involved research teams from several Italian universities: Padua, Chieti-Pescara, Turin, Bari, and Naples-Federico II.

The resilience test used in the survey is based on a selection from the 25-item Connor and Davidson (2003) resilience scale, as adapted by Campbell-Sills and Stein (2007) and translated into Italian by Bollani et al. (2023). The selected items were derived from a unidimensional factor analysis of the responses to the following 8-item question (Cascella et al. 2024): "On a 1 to 5 scale, where 1 means 'Not at all' and 5 means 'Very much', how true are the following statements for you?" The eight items are: a) I can deal with whatever comes; b) I try to see the humorous side of problems; c) Coping with stress can strengthen me; d) I tend to bounce back after illness or hardship; e) I can achieve goals despite obstacles; f) I can stay focused under pressure; g) I think of myself as a strong person, and h) I can face unpleasant or agonizing feelings.

Depression was measured using the 9-item PHQ – Patient Health Questionnaire, as proposed by Spitzer et al. (1999) and translated into Italian by Mazzotti et al. (2003). For this study, the measure was converted into a dichotomous variable, where 1 represents "severely depressed".

The other variables used in the statistical analysis are as follows.

Psychological capital. Variables include passive attitude (dichotomous), and proactive attitude (dichotomous), which were derived from a factor analysis of the responses to a set of eight items related to Future Time Perspective (FTP), selected from the 20 items of the Beck Hopelessness Scale Beck et al. (1974), as proposed by (Bollani et al., 2023). The first dichotomous variable was created from standardized scores below -0.25, and the second from scores above 0.40.

Social capital. Variables include Trust1, referring to people trust in health scientists and services (WHO, EU, hospitals, general practitioners, and local public bodies) and the information broadcast through television; Trust2, which reflects how people managed relationships with family, friends and neighbours during the pandemic (Bollani and Fabbris, 2024); Possession of a higher education degree; Cohabitation with a spouse; and Living in a family with children.

A complete list of the variables is presented in Table 1. The control variables (all dichotomous) are gender (male), age group 18-25, geographical area (Southern Italy) and student status. Initially, employee status was also included but was later excluded due to multicollinearity in the regression models.

2.2 Analytical models

Two distinct regression models were considered: one with resilience (RE) as the dependent variable, and the other with severe depression (SD) as the dependent variable. In both models, the other of the two variables is considered as a potential predictor, along with a selection of other relevant variables. Additionally, the same set of control variables was included in both models. The relationships can be expressed as follows:

$$RE = f(SD, X_{01}, \ldots, X_q \mid Z_1, \ldots, Z_4)$$
$$SD = f(RE, X_{01}, \ldots, X_q \mid Z_1, \ldots, Z_4).$$

The analytical model included two sets of regressors: the first set $Z_1, \ldots, Z_4$ of control variables, which were forced blockwise into the model, while the second set $X_{01}, \ldots, X_q$ ($q \leq 41$) includes variables selected for their statistical significance using a forward stepwise selection procedure.

The first model is an ordinary linear regression (with RE as a continuous dependent variable), whereas the second model, with SD as the dependent variable, is of a logistic regression. Logistic regression is used to predict binary outcomes: the dependent variable is a logarithmic transformation of the ratio of the probability of the event (SD = 1) to the probability of its complement (SD = 0). The dependent variable is modelled as a linear combination of regressors, and the coefficients of the regressors are referred to as β coefficients. Formally, the logistic regression is expressed as follows (Hosmer and Lemeshow, 2000):

$$logit\,[p(SD = 1)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$$

where $logit(p) = ln[(p/(1-p)]$, and β_i measures the relation between Y and the predictor X_i, while holding all other predictors constant. A potential regressor is included in the model only if it is

statistically significant. A relatively broad significance level ($p < 0.10$) was adopted for selecting the predictors.

Statistical analyses were performed using R (R Core Team, 2024). The *My.stepwise* package was used to select significant regressors via a forward stepwise procedure. Specifically, the *My.stepwise.lm* function was used for linear regression analysis, while the *My.stepwise.glm* function was used for generalised linear models (in this case, the logit model). These functions allowed us to set a ***s***ignificance ***l***evel for ***e***ntering the variables into the model (we used *sle* ≤ 0.10) and a ***s***ignificance ***l***evel for them to ***s***tay (we used *sls* = 1, consistent with a forward selection procedure). The *glm* package was used to compute confidence intervals for some of the regressors and the *DeskTools* package was employed for regression diagnostics.

3. Results

Our sample consists of 362 women (65.8% of the young adult sample). Among the respondents, 106 (19.3%) were employed and 429 (78%) were students. One-third of the young adults in the sample resided in the southern regions of Italy. The distributional parameters for severe depression among young adults, broken down by control variables, are presented in Table 1; the distribution of resilience scores is presented in Figure 1. The distributional parameters for the significant regressors are shown in Table 2.

Table 1. Descriptive statistics and relation between of the control variables and both RE-resilience and severe depression-SD (n = 550)

Variables	Sample %	Mean resilience	r(RE, x_r) + signific	% depressed	r(SD, x_r) + signific
Male	34.2	0.048	0.181 ***	17.0	-0.118 **
Age: 18-24	70.7	-0.284	-0.131 **	27.0	0.109 *
Southern Italy	33.3	0.037	0.170 ***	21.9	-0.035
Employed	19.3	0.101	0.149 ***	13.2	-0.123 **
Student	78.0	-0.281	-0.153 ***	26.8	0.124 **

*Significance: *** = < 0.001; ** = < 0.01; * = < 0.05; ° = < 0.10.*

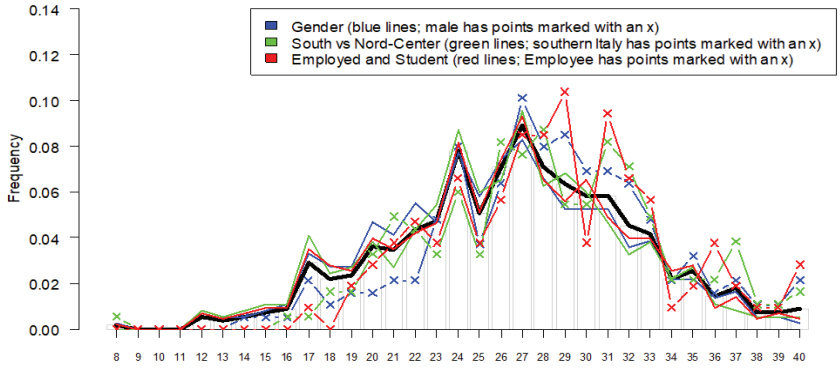


Figure 1. Frequency distribution of resilience scores for Italian young adults, by control variables

The results of the multivariate regression analysis with either the resilience and depression as criterion variables are presented in Table 3. Only the control variables and the significant predictors entering the model are presented. The results are presented in homogeneous blocks.

- The correlation coefficient between depression and resilience is -0.261. While the negative sign is expected, the magnitude of the correlation is smaller than expected. However, this

value increases when additional variables are included in the regression model. Specifically, the beta coefficient between severe depression and resilience increases to -0.431 when other behavioural and psychological variables are accounted for. This suggests that there are subgroups within the population where the relationship between depression risk and resilience is stronger than in the general population, and others where the relationship is weaker.

Table 2. Descriptive statistics and relation between significant regressors and either RE-resilience and severe depression-SD (n = 550)

Variables	Sample %	Mean resilience	r(RE, x_r) + signific	% depressed	r(SD, x_r) + signific
Physical damages from contagion	10.9	-0.260	-0.021	45.0	0.172 ***
Psychological damages from contagion	36.4	-0.472	-0.208 ***	41.0	0.301 ***
Consumed drinks more than before	8.7	-0.133	0.021	37.5	0.098 *
Consumed wine/beer more than before	9.5	0.033	0.076 °	32.7	0.066
Passive attitude (FTP—)	37.6	-0.595	-0.310 ***	43.0	0.346 ***
Proactive attitude (FTP+)	25.6	0.238	0.261 ***	6.40	-0.242 ***
Fear of Italian virus diffusion	94.7	-0.229	-0.120 **	24.6	0.056
Chronic disease	16.4	-0.268	-0.072 °	44.4	0.231 ***
Trust in health science	31.6	-0.027	0.119 **	19.0	-0.080 °
Trust in own networks	26.0	0.066	0.160 ***	18.9	-0.071 °
Doubting vaccines' effectiveness	11.8	-0.157	0.016	29.2	0.045
Distance learning during pandemic	18.0	-0.190	0.005	42.4	0.202 ***
Future income may be lower than before	32.9	0.011	0.150 ***	23.2	-0.013
Progress could stop after pandemic	7.8	-0.155	0.013	23.3	-0.005

*Significance: *** = < 0.001; ** = < 0.01; * = < 0.05; ° = < 0.10.*

- Men have a higher resilience score than women, holding all other predictors constant. Depression, which is slightly higher among women, is a significant predictor when considered in isolation, but when combined with other relevant predictors, it is no longer significant. Specifically, gender is not a significant predictor of depression when resilience is taken into account.
- Residents in Southern Italy exhibit significant higher resilience than those from other Italian regions, as expected, given the well-known coping strategies associated with Southern Italians. In contrast, there is no significant difference in depression rates between those living in the South and those from other regions in Italy.
- Two key findings emerge concerning health and disease. First, the pandemic has had a profoundly negative impact on psychological health, particularly in terms of depression. Second, chronic diseases are positively correlated with increased depression risk. The first result indicates that the pandemic triggered widespread fear of contagion, which exacerbated psychological distress, particularly among those who avoided hospitalisation. The second result shows that depression rates were notably higher among young adults with lower coping resources and then with a higher risk of severe consequences in case of contagion.
- Youth attitudes played a significant role in enhancing resilience, regardless of whether individuals had been infected by the virus. Variables that notably improved resilience included: the belief that health science could combat future viruses, the perception that a strong network of personal relationships would provide protection in the face of future crises, and the awareness that if similar events would happen again, people should rethink the social development pace, possibly at the detriment of their own social and economic status.
- The risk of severe depression is primarily associated with a lack of resilience, combined with other psychological vulnerabilities. Additional significant predictors of depression include

anti-vaccine beliefs and the experience of forced distance learning during the pandemic.

- Finally, some individuals with higher resilience scores reported consuming more wine or beer during the pandemic than before. In contrast, individuals with depressive tendencies stated they consumed more alcohol. While the latter finding aligns with expectations, as depressed individuals often turn to alcohol for relief, the former result is more surprising. Indeed, most young people reported no change in their drinking habits, and only a small proportion (9.5%) reported drinking more alcohol, compared to a larger group (19.4%) who reported drinking less. We did not find similar results in the literature.

Table 3. Beta estimates and their significance of the regression models with resilience and severe depression of Italian young adults as criterion variables (n = 550; forward stepwise selection criterion; p-value for variable selection: 0.10)

	Y=Resilience			Y=Severe depression		
	Beta estimate	Std. error	Signi-ficance[a]	Beta estimate	Std. error	Signi-ficance[a]
Intercept	-0.027	0.214		-2.906	0.443	***
Gender (Male)	0.289	0.082	***	-0.216	0.272	
Age class 18-24 (dichotomous)	0.034	0.100		0.327	0.337	
Southern *vs.* Other Italian regions	0.215	0.085	*	0.345	0.280	
Employed	0.026	0.135		-0.537	0.410	
Severe depression	-0.315	0.099[b]	**			
Resilience				-0.431	0.137[b]	**
Physical damages from contagion	0.238	0.127	°			
Psychological damages from contagion	-0.231	0.087	**	0.940	0.246	***
Consumed drinks more than before				1.014	0.394	*
Consumed wine/beer more than before	0.279	0.129	*			
Passive attitude (FTP—)	-0.351	0.091	***	1.216	0.270	***
Proactive attitude (FTP+)	0.241	0.098	*	-1.018	0.432	*
Fear of Italian virus diffusion	-0.412	0.171	*			
Chronic diseases				0.991	0.245	***
Trust in health science	0.204	0.081	*			
Trust in own networks	0.219	0.088	*			
Doubting vaccines' effectiveness				0.712	0.364	.
Distance learning during pandemic				1.040	0.300	***
Future income may be lower than before	0.222		*			
Progress could even stop after pandemic				-0.917	0.482	°
R^2	*0.243*		***	*0.393*		
F	*12.27*		***			

*(a) Significance: ***: < 0.001; **: < 0.01; *: < 0.05; °: < 0.10; a blank space points out a not significant estimate.*

(b) The 95% confidence interval for Severe depression is [-0.508;-0.122], that for Resilience is [-0.703;-0.166]

4. Discussion

This paper examined whether and how resilience and depression can be considered reciprocal conditions. Our findings, consistent with existing literature, show that a lack of resilience is significantly correlated with severe depression among Italian young adults. Specifically, we found that low levels of resilience, combined with a closed mindset towards the future, often predispose individuals to psychological distress. Conversely, high resilience, along with a proactive mindset, acts as a protective factor against severe psychological issues, particularly depression.

Mental difficulties in youth emerge more quickly and in a more severe form not only when

resilience and proactivity are weak, but also when individuals are ill-prepared to cope with life's challenges, are burdened by chronic illness, or perceive a lack of supportive social networks to shield them from external shocks, such as the COVID-19 pandemic. The interplay of any one of these factors with a lack of psychological capital during the pandemic allowed severe psychological distress to take hold, significantly increasing the risk of depression.

Finally, we found that, among young Italians, the fear of contracting the Coronavirus had a greater impact on psychological distress than the infection itself.

Our results suggest that, since social shocks cannot be avoided, the type and severity of post-traumatic symptoms should be carefully monitored by health and government authorities to mitigate long-term effects of crises on mental health, particularly among young people.

References

Barzilay, R., Moore, T.M., Greenberg, D.M., Didomenico, G.E., Brown, L.A., White, L.K., Gur, R.C., Gur, R.E. (2020). Resilience, COVID-19–related stress, anxiety and depression during the pandemic in a large population enriched for healthcare providers. *Translational Psychiatry*, 10(291): pp. 1–8.

Beck, A. T., Weissman, A., Lester, D., Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology*, 42(6): pp. 861–865.

Bollani, L., Di Zio, S., Fabbris, L. (2023). After-COVID remote working and towards-the-future disposition in Italy. In L. Fabbris, S. Mignani, G. Tassinari (eds). *Statistics, Technology and Data Science for Economic and Social Development. Book of Short Papers of the ASA Bologna Conference, Statistica Applicata – Italian Journal of Applied Statistics*, 35(3), pp. 57-62.

Bollani L., Fabbris, L. (2024). Trust in health science and organizations in Italy during the COVID-19 pandemic (submitted).

Brooks, S.K., Webster, R.K., Smith, L.E., Woodland, L., Wessely, S., Greenberg, N., Rubin, G.J. (2020). The psychological impact of quarantine and how to reduce it: Rapid review of the evidence. *The Lancet*, 395(10227), pp. 912-920.

Campbell-Sills, L., Stein, M.B. (2007). Psychometric analysis and refinement of the Connor-Davidson Resilience Scale (CD-RISC): Validation of a 10 item measure of resilience. *Journal of Traumatic Stress*, 20(6): pp. 1019-1028.

Cascella, C., Bollani, L., Fabbris, L. (2024). Measuring persons' resilience: Properties of an Italian inventory scale (submitted).

Connor, K.M., Davidson, R.T. (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC), *Depression and Anxiety*, 18: pp. 76–82.

Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression (2nd ed.)*. New York: John Wiley & Sons.

Luthar, S.S., Cicchetti, D. (2000). The construct of resilience: Implications for interventions and social policies. *Development and Psychopathology*, 12(4): pp. 857–885.

Mazzotti, E., Fassone, G., Picardi, A., Sagoni, E., Ramieri, L., Lega, I., Camaioni, D., Abeni, D., Pasquini, P. (2003). II Patient Health Questionnaire (PHQ) per lo screening dei disturbi psichiatrici: Uno studio di validazione nei confronti della Intervista Clinica Strutturata per il DSM-IV asse I (SCID-I). *Italian Journal of Psychopathology,* 9(3): pp. 235-242.

Preston, A., Rew, L., Calloway Young, C. (2023). A systematic scoping review of psychological capital related to mental health in youth. *The Journal of School Nursing*, 39(1): pp. 72–86.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AT (http:/www.R-project.org,, last access 10 July 2024).

Spitzer, R.L., Kroenke, K., Williams, J.B. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ Primary Care study. *Journal of the American Medical Association*, 282(18): pp. 1737-1744.

doi.org/10.26398/asaproc.0081

Trust in health science and organizations in Italy during the Covid-19 pandemic

Luigi Bollani[1], Luigi Fabbris[2]
[1] Dept. of Economics, Social Studies, Applied Mathematics & Statistics, Univ. of Turin, Italy
[2] Tolomeo Studi e Ricerche, Padua, Italy

1. Introduction

In this paper, we deal with social trust, which refers to the belief in the integrity and reliability of a social body. In the context of health science, people's trust centres on the concept of scientific integrity of scientific statements (ten Have and Gordijn, 2018). With reference to healthcare organisations, trust concerns their ethical governance, transparent health policies, and the effective provision of services to the public (Taylor et al., 2023b).

Trust in health science was particularly relevant during the Covid-19 pandemic, given the measures imposed to contain the virus spread. National governments worldwide introduced lockdowns and other strict behavioural rules lasting for months, while vaccination campaigns reached virtually all people. The reduction of the coronavirus in 2022 was expected to mark the peak of public trust in both health science and healthcare organisations. However, scholars and observers generally agree that the end of the pandemic coincided with a decline in trust in science and healthcare institutions. Specifically, the World Economic Forum (2023) argues that, in the US, there was a significant lack of trust in medical scientists to act in the public's best interests. The pandemic further diminished the already modest levels of trust in many, if not all, of the key actors involved in healthcare delivery (Taylor et al., 2023a).

Undoubtedly, the Covid-19 pandemic was a period during which trust in health science and healthcare institutions was brought into sharp focus. In fact, scholars argued that the pandemic presented an opportunity for telemedicine to become more widely accepted, for information networks and collaboration around healthcare to expand across communities, and for the personalisation of medical information. Rosen et al. (2020) suggested that the pandemic showed the potential of remote consultations, which—during times of contagion— eliminate the risk of transmission in crowded waiting rooms, and, in normal times, can occur without necessarily compromising the quality of care provided by healthcare professionals.

Regarding vaccinations, a minority of the population opposed them. This group, known as the 'no-vax' movement, was particularly vocal during and after the pandemic, questioning the effectiveness of the vaccines and claiming that vaccination posed risks to people's health (Daly et al., 2020; Sturgis et al., 2021). Scepticism about vaccination also extended to scientists, who—after initial enthusiasm for the newly developed vaccines—highlighted the need for monitoring their effects before proceeding with a third and then a fourth vaccination campaign. In fact, participation in the third and fourth campaigns was so low compared to the first two that millions of vaccine doses prepared for the fourth campaign were left unused and subsequently destroyed. It is possible that this hesitancy reflects a lack of confidence in "those in charge of administering" medicine (Rand et al., 2023).

To explore Italians' trust in health science and organisations, we conducted a survey of Italian adults using a computer-assisted web interviewing (CAWI) questionnaire. This survey formed part of the RECALL project, which aimed to study the well-being of Italian adults in the aftermath of the pandemic (Bollani et al., 2023). In the following sections, social trust is treated as an independent variable related to attitudinal and behavioural outcomes.

Our research questions are as follows:
Q1: What is the level of social trust among Italian adults in the aftermath of the Covid-19 pandemic?

Q2: Are there population groups that place greater trust in health science and services, and conversely, are there groups that are more sceptical towards health institutions? Q3: Is Italians' social trust related to their own or their families' experience of infection?

The remainder of the paper is structured as follows: Section 2 introduces the available data and the model for data analysis. Section 3 presents the main results of the statistical analysis. Finally, Section 4 discusses the results in relation to the existing literature.

2. Data and methods

An extensive questionnaire was administered using the CAWI technique from June to November 2021 by research teams from various Italian universities (Padua, Chieti-Pescara, Turin, Bari, and Naples). We obtained 1,192 responses, which, after checking for nonresponse to the question about resilience, were reduced to 1,182 valid questionnaires. Excluding 7 respondents residing abroad and another who did not provide gender information, we were left with 1,174 valid responses from Italian adults.

The question on social trust during the pandemic was: "*On a scale from 1 to 5, where 1 is the minimum and 5 is the maximum, how much do you agree with the following statements?*" Eleven statements followed: 1. During the pandemic, the World Health Organization provided the right guidance; 2. My respect for the European Union increased after it dealt with the emergency; 3. People were more likely to be infected within hospitals than outside them; 4. Public bodies (municipalities, health units, etc.) managed the epidemic promptly and efficiently; 5. There is no doubt that scientists contributed to fighting the virus; 6. I admired the availability of general practitioners during the epidemic; 7. I was correctly informed by the media during the epidemic; 8. Social media were a valuable, albeit unofficial, source of information about the virus; 9. During lockdowns (home confinement), I was reassured by the presence of local shops; 10. During the epidemic, friends were of significant help to me; 11. During the epidemic, I received considerable support from my relatives.

The responses to the eleven trust-related questions were factor analysed and the loadings were Varimax rotated (Table 1). The analysis revealed two significant factors: **Trust1**, which involves appreciation for health scientists, international and local public bodies, and the media and may be termed "trust in health science and organisations", and **Trust2**, which involves the set of variables describing individuals' informational and relational networks, may be referred to as "trust in personal networks". The two factors are linearly independent and therefore represent two additive phenomena. This suggests that, in responding to the pandemic, Italians created a network of relationships around them as an additional—and in some ways alternative—tool to the efforts of the public bodies in charge of fighting the virus.

Due to space limitations, we present only the variables that were significant in the multivariate model.

- **Covid-19**: Scoring 0 if the respondent was infection-free (no positive test) throughout the pandemic, and 1 for those who were infected; and **Damage_Psy**: For respondents who reported psychological damage following the pandemic experience.
- **FTP –** Future Time Perspective: Derived from a factor analysis of responses to a set of eight items related to pessimism-proactivity, selected from the 20 items of the Hopelessness Scale proposed by Beck et al. (1974) and translated into Italian by Bollani et al. (2023). From the standardised factor scores, we created two dichotomous variables: one indicating a **passive attitude** (below -0.25 standardised scores) and another for a **proactive attitude** (above 0.40 standardised scores).
- **Resilience**: Derived from factor analysis of responses to eight questions concerning a resilience scale, translated into Italian by the authors (Bollani et al., 2024). From the standardised factor scores, two dichotomous variables were created: **HighR**, for scores at or above the seventh decile, and **LowR**, for those below the third decile.
- **Spouse** (living with a spouse) and **Children** (living in a household with children), as

indicators of social capital.

- **Chronic**: 0 for no chronic disease, 1 for one chronic disease, and 2 for two or more chronic diseases; and **Psychic**: Indicates individuals affected by chronic psychological disorders.
- **JobLoss** (loss of employment); **RW** (working remotely); **DT** (learning or teaching remotely); and **Sports** (engaging in sports more than before).
- **OwnRisk**: Perception of infection risk for oneself; **ItalyRisk**: Perception of infection risk for Italians in general; **DoubtVaccine** (doubts about vaccine effectiveness); **FearVaccine** (fear of vaccine side effects); **RWJudge** (evaluation of remote working); **DTJudge** (evaluation of distance learning); **PoorFuture** (no concern about being poorer in the future); and **NoProgress** (no need for future social progress).

The analytical model included two sets of predictors. The first set, $Z_{01}, \ldots, Z_{05}$, comprised control variables that were block-wise forced into the model, except for "Employed" and "Student," which were initially included but later excluded due to their inflationary effects on estimation variance. The second set, $X_{01}, \ldots, X_{21}$, was selected based on statistical significance using a forward stepwise selection procedure.

The relationship between Y and the predictors can be expressed as:

$$Y=f(X_{01},\ldots, X_{21} \mid Z_{01},\ldots, Z_{05}).$$

All statistical analyses were performed using R (R Core Team, 2024), applying the *My.stepwise* procedure to select the significant predictors in the regression analysis. A significance level of < 0.10 was adopted for predictor selection.

Table 1. Distributional statistics and results of the Varimax-rotated factor solution of social entities upon which people relied during the Covid-19 pandemic (n = 1,174).

Social entity	*Mean*	% ≥ 4	*Factor 1*	*Factor 2*	*Communality*
1: WHO	2.84	25.0	0.780	0.058	0.612
2: EU	2.76	21.3	0.727	0.059	0.532
3: Hospitals	2.46	15.3	-0.198	0.519	0.309
4: Local public bodies	2.69	21.8	0.643	0.119	0.428
5: Scientists	3.93	68.9	0.707	-0.030	0.501
6: General practitioners	3.13	40.5	0.583	0.128	0.356
7: TVs	2.29	13.6	0.576	0.196	0.370
8: Social media	2.38	19.8	0.113	0.400	0.173
9: Neighbouring shops	3.27	42.6	0.222	0.594	0.408
10: Friends	2.87	27.6	0.172	0.472	0.252
11: Relatives	3.59	55.7	0.035	0.638	0.408

3. Results

Table 1 shows that the vast majority of respondents endorsed scientists' ability to combat the virus and communicate their knowledge to a broad audience (68.9% endorsement). Furthermore, respondents reported receiving significant support during the pandemic from relatives (55.7%), neighbouring shops (42.6%), and general practitioners (40.5%). The lowest levels of endorsement were observed for TVs (13.6%), social media (19.8%), the European Union (21.3%), public bodies (21.8%), and the World Health Organization (WHO) (25.0%).

It is noteworthy that media endorsement was particularly low, with only 13.6% of respondents supporting TV channels and 19.8% endorsing social media. These media outlets played different informational roles: TV channels were associated with the health system's efforts to inform the public about the risks of contagion and prevention measures, while social media provided supplementary information on the risk of contagion and suggested therapies in case of infection. It is possible that some respondents viewed social media as a counter-informational platform in contrast to TV, which they may have considered the "official" voice.

The very low communality found in the factor analysis indicated that the question about infection risks within hospitals did not strongly belong to either of the two factors. Although the question on in-hospital infections was intended to gauge respondents' "no-vax" tendencies, it appears that respondents interpreted it differently. This prompted us to analyse the responses to this question separately. The results of the multivariate regression analysis conducted on the two factor scores are summarised in Table 2. It appears that the predictors explain a fair amount of the variability of the first factor ($R^2 = 0.258$), while explaining a smaller portion of the variability of the second factor ($R^2 = 0.129$).

Table 2. Beta estimates and significance of the linear regression model with criterion variables: trust in health science and organisations (Trust1) and trust in personal networks (Trust2) of Italian adults (n = 1,174; stepwise selection criterion: forward; p-value for variable selection: 0.10).

	Trust 1		Trust 2	
	Beta estimation	Sign.	Beta estimation	Sign.
Intercept	-0.098		-0.661	***
Gender (Male)	-0.077		-0.189	**
Age (years)[a]	-0.001		0.010	***
Southern vs. Other Italian regions	-0.150	**	0.135	*
Doubt vaccine effectiveness	-0.888	***		
Vaccine side effects fear	-0.547	***		
Distance learning or teaching			-0.255	***
Distance learning positive judgement	0.245	***		
Worked from remote			-0.227	**
Remote working positive judgement	0.145	*		
High resilience score	0.161	**	0.315	***
Low resilience score	-0.124	*	-0.163	*
Fear own infection	0.128	*	0.145	*
Fear Italy infection	0.314	*	0.466	***
Impact: psychological damages	-0.147	*		
Chronic diseases			-0.182	***
Psychic chronic disease	0.363	*		
No problem to become poorer in the future	-0.146	*		
Passive personality			-0.105	.
With spouse	0.184	°		
With children			-0.185	.
Covid-19 infected			0.255	***
No need social progress			0.300	**
Own income lower in the future			0.166	*
Sports practicing			-0.103	°
R^2	*0.258*	***	*0.129*	***
F	*26.69*	***	*10.75*	***

(a)The lower age class was defined as 18-24 in the analysis of young adults and 18-34 in the analysis of the full sample.
*Significance: ***: < 0.001; **: < 0.01; *: < 0.05; °: < 0.10; a blank space indicates a non-significant estimate.*

The science trust factor (Trust1) correlates positively with being part of a couple residing in a northern or central Italian region, fearing viral contagion both for themselves and others, positively evaluating both remote working and learning, scoring high on resilience, and having a history of chronic psychological disorders. The correlation coefficient was negative for those with a "no-vax" attitude, lower resilience, perceived psychological damage, and a belief that

becoming poorer in the future would not be a problem. Trust in health scientists and institutions appears to be associated with a strong belief in science and related social bodies, while the opposite attitudes were expressed by people who were sceptical of official science and vaccines, those with psychological vulnerabilities prior to the pandemic, or those perceiving a bleak future.

The network trust factor (Trust2) correlates positively with being female, being older, residing in a southern Italian region, being highly fearful of infection risks and/or having been infected by the virus, having a proactive personality, being sceptical about future social progress, and being prepared for a potential reduction in income in the future. Characteristics negatively correlated with the network factor include not engaging in remote working or learning, not having children, not practising sports, and having low or no resilience.

4. Discussion and conclusion

The global health crisis of the COVID-19 pandemic abruptly forced people to make decisions quickly, often without adequate information. People knew little about the virus and were compelled to communicate, work, and make decisions with partial and often contradictory information. Trust in both official and unofficial sources of information became central. Our examination of Italians' confidence in scientists, health organisations, and support networks revealed a strong trust in science, a reliance on relational networks that some groups of people used to navigate their situations, and a rather tenuous alignment with the Italian or international bodies designated to manage the epidemic. This aligns with existing literature, which suggests that people tend to evaluate scientific institutions separately from scientific principles and tools. In general, they trust science but are sceptical of institutions (see, for example, Huber et al., 2019). Moreover, after a natural disaster—and the pandemic can be seen as such—surveys often show increased distrust towards authorities due to perceived shortcomings in their response (Rand et al., 2023).

Social communication deserves special attention. During the pandemic, people turned to both official (TV) and unofficial (social media) sources for guidance on appropriate behaviour. We found that they mistrusted both TV and social media, but they considered the latter a more reliable part of their support system, despite it being sometimes openly opposite to official communication. Similar findings were reported in the US by Huber et al. (2019), who argued that people trusted science news on social media because it came from social contacts they trusted.

Our findings also identify the core of people's trust in the support they receive from relatives, friends, and local communities. During the pandemic, families acted as the social microcosm that helped members organise their lives and foster resilience. Sharing meals, rediscovering old games, reorganising sleep schedules, redistributing care responsibilities, and competing for home spaces for remote working or learning became part of the "new normal" (Rania et al., 2021; Gayatri and Puspitasari, 2022; Partington et al., 2022). However, research on psychosocial issues has also highlighted an increase in domestic conflicts, violence, and caregiving burdens during the pandemic (Soejima, 2021), though these negative phenomena were limited to a small minority of Italian families.

People also attributed a role to local shops, ranking them just below relatives in importance. During lockdowns, these shops became community meeting points, akin to a modern agora. People could enhance their knowledge of organic food, health products, and the logistics of hoarding and planning family needs. Moll-Brandão et al. (2023) argue that the pandemic transformed shops from simple retail outlets to influential spaces where neighbours and experts could meet. Whether this shift will persist is unclear, but the pandemic experience may offer valuable insights for shaping the future of retail and for repositioning shops in city centres. We can conclude that, despite its highly negative health and social impacts, the pandemic also prompted positive and beneficial changes in people’s and families’ lives.

References

Bollani, L., Di Zio, S., Fabbris, L. (2023). After-COVID remote working and towards-the-future disposition in Italy. In L. Fabbris, S. Mignani, G. Tassinari (eds). *Statistics, Technology and Data Science for Economic and Social Development. Book of Short Papers of the ASA Bologna Conference, Statistica Applicata – Italian Journal of Applied Statistics*, 35(3), pp. 57-62.

Daly, M., Jones, A., Robinson, E. (2020). Public trust and willingness to vaccinate against COVID–19 in the US from October 14, 2020, to March 29, 2021. *Journal of the American Medical Association*, 325(23), pp. 2397-2399.

Gayatri, M., Puspitasari, M.D. (2022). The impact of COVID-19 pandemic on family well-being: A literature review. *The Family Journal: Counseling and Therapy for Couples and Families*, pp. 1-8. DOI: 10.1177/10664807221131006

Huber, B., Barnidge, M., de Zúñiga, H.G., Liu, J. (2019). Fostering public trust in science: The role of social media. *Public Understanding of Science*, 28, pp. 759–777.

Moll Brandão, M., França Sarcinelli, A., Bisi Barcelos, A., Postay Cordeiro, L. (2023). Leisure or work? Shopping behavior in neighborhood stores in a pandemic context", *RAUSP Management Journal*, 58(2), pp. 97-110.

Partington, L.C., Mashash, M., Hastings, P.D. (2022). Family thriving during COVID-19 and the benefits for children's well-being. *Frontiers in Psychology*, 13, 879195. doi: 10.3389/fpsyg.2022.879195

Rand, L., Carpenter, D.P., Kesselheim, A.S., Bhaskar, A., Darrow, J.J., Feldman, W.B. (2023). Securing the trustworthiness of the FDA to build public trust in vaccines, in L.A. Taylor, G.E. Kaebnick, M.Z. Solomon (eds.) *Time to Rebuild: Essays on Trust in Health Care and Science*, Special Report, Hastings Center Report 53/5, pp. S60-S68.

Rania, N., Coppola, I., Lagomarsino, F., Parisi, R. (2021). Family well-being during the COVID-19 lockdown in Italy: Gender differences and solidarity networks of care. *Child & Family Social Work*. 27, pp. 1–10.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AT (http:/www.R-project.org , last access 10 Sept. 2024).

Rosen, C.B., Joffe, S., Kelz, R.R., MSCE, MBA (2020). COVID-19 moves medicine into a virtual space. A paradigm shift from touch to talk to establish trust. *Annals of Surgery*, 272(2), pp. e159-e160.

Soejima, T. (2021). Impact of the coronavirus pandemic on family well-being: A rapid and scoping review. *Open Journal of Nursing*, 11, pp. 1064-1085.

Sturgis, P., Brunton-Smith, I., Jackson, J. (2021). Trust in science, social consensus and vaccine. *Nature Human Behaviour*, https://doi.org/10.1038/s41562-021-01115-7.

Taylor, L., Nong, P., Platt, J. (2023a). Fifty years of trust research in health care: A synthetic review. *The Milbank Quarterly*, 101(1), pp. 126-178.

Taylor, L., Solomon, M.Z., Kaebnick, G.E. (2023b). Trust in healthcare and science: Toward common ground on key concepts. In L.A. Taylor, G.E. Kaebnick and M.Z. Solomon (eds.) *Time to Rebuild: Essays on Trust in Health Care and Science*, Special Report, Hastings Center Report 53/5, S2-S8. DOI: 10.1002/hast.1517.

ten Have, H., Gordijn, B. (2018). Trust in healthcare and science. *Medicine, Health Care and Philosophy*, 21, pp. 157–158 (https://doi.org/10.1007/s11019-018-9840-3).

World Economic Forum (2023). Why has trust in science and healthcare dropped and how can we reverse this trend? *World Economic Forum Annual Meeting*. https://www.weforum.org/agenda/2023/12/trust-in-science-and-healthcare-drops-to-new-lows-wef24/

doi.org/10.26398/asaproc.0082

Artificial intelligence and future scenarios: Preliminary results from an application on family policies in north-eastern Italy

Mario Bolzan[a], Yuri Calleo[b], Simone Di Zio[c], Marco Marozzi[d], Manuela Scioni[a]
[a] Department of Statistical Sciences, University of Padova, Padova, Italy.
[b] Spatial Dynamics Lab, University College Dublin, Ireland.
[c] Department of Legal and Social Sciences, University "G. d'Annunzio" of Chieti-Pescara, Italy
[d] Department of Mathematics and Informatics, University of Ferrara, Italy

1. Introduction

In an era of great uncertainty, many public authorities are prioritizing the development of policies focused on family support. These initiatives aim to assess strategies that enhance family well-being and mitigate potential challenges. By investing in family-friendly policies, such as affordable childcare, parental leave, and financial support, national and regional entities aim to promote positive outcomes and reduce risks associated with economic and social pressures. These efforts are important in creating resilient family structures and fostering a positive environment for future generations. In this paper, we aim to contribute to research on *anticipatory governance*, a well-established approach that focuses on imagining, exploring, and preparing for future challenges and changes before they fully emerge (Fuerth, 2009). Unlike traditional reactive governance – which addresses problems after they arise – anticipatory government seeks to anticipate potential issues and implement proactive measures to address them. Different key aspects characterize this approach, including foresight and scenario planning (Calleo and Pilla, 2023), adaptive policies, innovation and technology, collaboration and engagement, and resilience building. We are currently in what is defined as the "age of data", where the proliferation of big data and advancements in artificial intelligence (AI) make data crucial for developing anticipatory solutions (Di Zio et al. 2023). These technologies offer a deeper, more detailed understanding of potential future developments with increased precision and speed. This integration of advanced data analytics into governance processes not only improves the accuracy of actions but also supports more strategic and adaptive policymaking, ultimately contributing to more resilient and forward-thinking governance. In addition, there is considerable interest in exploring the "scenarios" that institutions, perhaps unconsciously, through their promotion of policies, envision for the future, and compare them with the viewpoints of experts.

These considerations are particularly relevant in the realm of family policies, a constantly evolving and complex area exposed to rapid changes in social dynamics, economic conditions, and demographic trends (Bolzan, 2018). Anticipatory governance plays a crucial role in shaping effective family policies by enabling policymakers to proactively address emerging challenges and opportunities. As family dynamics and structures evolve, it is essential for policies to be designed with an eye toward future trends and plausible scenarios rather than simply reacting to current issues. For example, shifts in demographic patterns, such as declining birth rates or an ageing population, necessitate policies that can adapt to these changes before they create significant challenges. Economic factors also illustrate the importance of anticipatory governance. Economic downturns or shifts in the job market can have profound effects on family stability and well-being. By anticipating these economic trends, policymakers can develop measures to support families during periods of financial instability, such as by enhancing unemployment benefits or creating job training programs. Technological advancements present another area where anticipatory

governance is vital. As technology continues to evolve, it affects family life in various ways, from changing work environments to altering communication patterns and parenting practices. Policymakers who anticipate these changes can better prepare family support systems to remain relevant and effective. Social changes are also a key consideration. Evolving attitudes toward issues such as gender roles, same-sex marriage, and parental leave require policies that are not only current but also forward-thinking. In essence, anticipatory governance helps ensure that family policies reflect and support these changing social norms, promoting inclusivity and equity.

This paper examines the legal initiatives undertaken and implemented by the administrations of the four regions of the Northeast of Italy: Friuli Venezia Giulia, Trentino-Alto Adige, Veneto, and Emilia Romagna during the four years from 2018 to 2022, to extract from them the advised scenarios for the family's future and compare them with those obtained from the experts.

Nevertheless, analyzing multiple documents, whether manually or through text-mining techniques, may provide only a partial achievement of the goal due to the lack of a clear understanding of the issue and the difficulty of extraction. To overcome these challenges, the research objectives (RO) of this paper are the following:

- RO_1: Identify and extract regional legal initiatives: investigate and extract legal initiatives from selected regions in Northeastern Italy and analyze these initiatives using artificial intelligence models.
- RO_2: Analyse regional family issues: employ Generative Pre-Trained Transformers to identify and outline common issues and needs experienced by families within the region.
- RO_3: Develop policy recommendations: utilize Generative Pre-Trained Transformers to generate a list of potential policy recommendations tailored to each region, which can be presented to policymakers for further discussion and refinement.

The rest of the paper is organized as follows: Section 2 details the methodology adopted, Section 3 presents and discusses the results, and Section 4 concludes.

2. Materials and methods

The methodology adopted for this paper started from the extraction of legal initiatives from the official websites of the four regions analyzed: Friuli Venezia Giulia, Trentino-Alto Adige, Veneto, and Emilia-Romagna in the four years (2018-2022). The extraction is performed using either each site's internal search engine or the advanced features of the Google search engine. All documents related to regional legal initiatives containing the word "family" or similar terms were extracted. The documents thus obtained were analyzed one by one to eliminate any duplicate or irrelevant documents. The extraction allows us to obtain a data matrix containing all the relevant information to import into the model, including the region of reference, type of document, associated link, associated administrative number, date, title, and text. To facilitate semi-automated analysis of the matrix, we employed ChatGPT-4, an advanced generative AI model. Developed by OpenAI, ChatGPT-4 (www.chatgpt.com) is an AI language model engineered to comprehend and generate text that mimics human language. This model utilizes a transformer-based machine learning architecture, which enables it to effectively process and produce natural language outputs based on the provided input. The matrix is subsequently converted to Portable Document Format (PDF) to facilitate uploading. Four primary prompts were then utilized within the document for a deeper analysis. The selection of prompts is inherently subjective and has a substantial impact on the final outcomes. Nonetheless, we opted to use general prompts with straightforward formulations, defined as follows:

- **P1**: Act like a policymaker. Considering these official documents, provide a list of policies that support families in the region [*insert region here*], Italy.
- **P1.1**: Considering this list of family support policies of the region [*insert region here*],

Italy, eliminates those that are not relevant.

- **P2**: Act like a family member. Considering these official documents, what issues and needs emerge for families in the [*insert region here*] region in the text?
- **P2.1**: Considering this list of emerging needs of the region [*insert region here*], Italy, eliminates those that are not relevant.

From a statistical point of view, when we import the document into the model, asking a specific question, the initial step involves analyzing the content of the document employing natural language processing (NLP) models to understand the document's content, including 1) tokenization; 2) entity recognition; and 3) structural analysis. In our case, since we have a document D for each region, consisting of words $w_1, w_2, \dots, w_n$, the process involves breaking D into tokens $t_1, t_2, \dots, t_m$, where each token t_i, can be a word or a sub-word. Once this process is concluded, the entity recognition involves identifying entities E from the tokens adopting a named entity recognition (NER) (Wang et al. 2023), thus allowing for the extraction of relevant information crucial for answering the specific prompt considering the documents. In this case, if t_i corresponds to a location, E_i might be tagged as a location entity, using a conditional random field (CRF) model or transformers. Finally, the structure of the text is analysed adopting a probabilistic parsing, where given a sequence of tokens $(t_1, t_2, \dots, t_m)$, a parse tree P is generated to represent syntactic structures. This can be denoted as $P = argmax_p \Pr(P|t_1, t_2, \dots, t_m)$, where $\Pr(P|t_1, t_2, \dots, t_m)$ represents the probability of parse tree P given the token sequence.

Specifically for our questions, with regards to P1 and P2, the model looks for relevant sections in D. Let Q be the query (the prompt selected), and S be the section of D. The relevance score $R(Q,S)$ is computed, adopting a cosine similarity (Ethayarajh, 2019) in the embedding space defined as:

$$R(Q,2) = \frac{\boldsymbol{E}_Q \cdot \boldsymbol{E}_S}{||\boldsymbol{E}_Q||\,||\boldsymbol{E}_S||}, \tag{1}$$

where $\boldsymbol{E}_Q$ and $\boldsymbol{E}_S$ are the embedding of the query and the section, respectively. For each relevant section S, policies P_i are extracted based on key phrases and entities. If S contains phrases associated with policies, the model uses learned patterns to identify these, utilising a classification approach defined as:

$$\hat{P} = argmax_p \Pr(P_i|S). \tag{2}$$

On the other hand, with regards to P2, if the model identifies family issues by looking for specific patterns or entities related to problems within D, this is denoted as $I = \{i_j \mid \Pr(i_j \mid t_1, t_2, \dots, t_m > \theta\}$. Where i_j are issues and θ is a threshold for filtering. Finally, to summarise the issues and needs, we adopt an extractive summarisation technique where the score for each sentence s in D is calculated as: $S(s) = f(s, I)$, where f is a function that scores the relevance of the sentence s to the issues I. Since, within the extracted policies some of them were on other topics (e.g., related to climate change), we decided to use an exclusion criterion by providing the GPT with prompt 1.1 and 2.1. In this case, once we have a list of policies ($L = \{p_1, p_2, \dots, p_n\}$) generated from the initial analysis, the model automatically uses proper keywords and key phrases to identify the main terms or phrases that describe the focus of each policy and the relevance criteria and contextual embeddings to represent policies and relevance criteria in a high-dimensional space (Ha et al. 2023). At the end of the process, we obtain a list L of policies and issues emerged from the four regions in short time frames ready for evaluation.

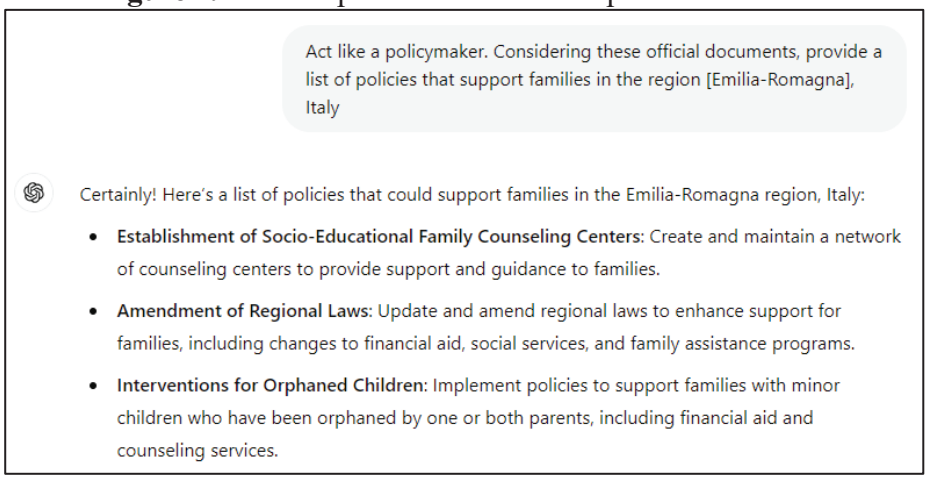
3. Results and discussion

The GPT-4 model was utilized to efficiently analyze administrative documents extracted from

the four regions. This approach enabled the extraction of a comprehensive list of intervention policies and provided a deeper understanding of the emerging needs highlighted within the documents. These results represent a dual advancement in the scientific literature: 1) First, they demonstrate the ability to rapidly extract information from a corpus of documents – within a matter of seconds – without requiring interpretative efforts from the analysts, as is often necessary with techniques like topic modelling. 2) Second, the approach enables the identification of actionable insights, in this case, formulating response policies for families in various regions based on the extracted information. Specifically, the model generated $L = 50$ preliminary policies for each region using the P1 prompt. Nevertheless, one limitation was the inclusion of policies unrelated to the family theme, as the model, after identifying a certain number of specific reference policies, continued to generate additional policies that were similar but not directly relevant. This issue arose from the model's tendency to extend the search beyond the essential reference policies. To address this issue, we employed the refined prompt 1.1, which instructed the model to eliminate non-relevant policies. As a result, the model produced 16 relevant policies for Emilia Romagna, 41 for Friuli Venezia Giulia, 47 for Trentino Alto Adige, and 47 for the Veneto region.

Considering the P2 prompt, the model efficiently identified the needs of families as expressed in the analyzed texts, extracting a total of 20 distinct needs across the four regions. In this instance, the results appeared consistent with the family theme and the prompt used. However, to ensure accuracy, we applied the refined prompt P2.1, allowing the model to further assess the relevance of the identified needs. The refinement process confirmed the relevance of all 20 needs extracted for each region, with no irrelevant needs identified. An example of the model's output is illustrated in Figure 1.

Figure 1. An example of the model's output.



The responses generated by the AI application are inherently influenced by the diverse formal languages employed by regional administrations, which reflect varying cultural perspectives and sensitivities regarding family structures and the social and economic challenges unique to each territory. Across all regions, economic assistance to families is prominently featured, particularly targeting those facing significant hardships such as large families, single parents, and individuals with disabilities. Additionally, there is a consistent emphasis on enhancing and expanding access to social and health services. However, notable differences between the four regions warrant further investigation. Below are specific elements observed in each region that were not explicitly present in the others. The Friuli Venezia Giulia region adopts a multifaceted approach to addressing several key social issues. Firstly, it places significant emphasis on combating school dropout rates and promoting academic success. This includes targeted programs designed to prevent bullying and address educational poverty, thereby supporting students who are at risk of falling behind. In addition to educational initiatives, the region is active in developing and managing social tourism activities. These programs aim to offer affordable travel opportunities for families, enhancing their

quality of life and providing valuable recreational experiences. Moreover, the Friuli Venezia Giulia region implements comprehensive strategies to tackle domestic violence. This includes providing support for victims and their families and ensuring that those affected by such violence receive the necessary help and resources to rebuild their lives.

In the Veneto region, several distinctive programs address specific family and social issues. The region places a strong emphasis on substance abuse prevention through targeted interventions aimed at reducing addiction within families and supporting affected individuals. Additionally, Veneto recognizes the need for legal assistance for families facing issues such as domestic violence, child custody disputes, or immigration concerns, and provides relevant support services. The region also prioritizes digital inclusion by ensuring that all families have access to technology and digital literacy training, which is essential for full participation in the modern digital environment. Furthermore, Veneto promotes inter-generational activities designed to strengthen relationships between different age groups, fostering greater cohesion within families and communities.

In the Trentino Alto Adige region, the approach to supporting families includes a focus on substance abuse and addiction services. The region offers a range of support and treatment options designed to assist families affected by these issues, aiming to mitigate the impact of substance abuse and provide pathways to recovery. Moreover, Trentino Alto Adige invests in cultural and recreational programs that enhance family life and foster community cohesion. These initiatives are intended to enrich the lives of families through diverse cultural experiences and recreational activities, strengthening social bonds and promoting a vibrant community environment.

In the Emilia Romagna region, several critical issues impact families in the region. One prominent challenge is digital inclusion; many families face barriers to accessing high-speed internet and digital devices, which limits their opportunities for online education, remote work, and participation in social activities. Environmental health also poses a significant concern, as poor environmental conditions, including pollution and inadequate waste management, adversely affect the well-being and quality of life of families in the region. Furthermore, the region struggles with underdeveloped transportation infrastructure. The public transportation network's limitations make it difficult for families to commute efficiently for work, school, and other essential activities.

These findings highlight areas where the AI approach is effective while also underscoring the need for continued experimentation with new methodologies and analytical tools to address these complex issues. It is important to acknowledge the role of regional socio-economic and cultural factors in shaping the outcomes of family policies. The divergence in policy effectiveness across regions such as Friuli Venezia Giulia, Trentino-Alto Adige, Veneto, and Emilia Romagna suggests that local contexts significantly influence how policies are perceived and implemented. For instance, the success of educational and social tourism initiatives in Friuli Venezia Giulia can be attributed to the region's strong emphasis on social cohesion and community-based interventions. Similarly, Veneto's focus on legal assistance and digital inclusion highlights the region's proactive stance in addressing modern societal challenges. However, these differences also underscore the limitations of a one-size-fits-all approach in policy development. The effectiveness of family policies is not solely dependent on the policies themselves but also on the underlying socio-economic fabric and cultural attitudes within each region. This nuanced understanding of regional dynamics should be a focal point for future policy formulation, ensuring that strategies are tailored to the specific needs and contexts of each area.

4. Conclusions

Anticipatory governance, as explored in this study, is a multifaceted approach that integrates statistical analysis, sociological insights, and artificial intelligence (AI) to proactively address future challenges in family policies. This paper investigated the application of anticipatory governance within the context of family policies in Northeastern Italy, considering advanced AI models, specifically GPT-4, to facilitate the extraction, analysis, and proposal of relevant policies across

four distinct regions. The research underscored the efficacy of AI-driven natural language processing techniques in rapidly analyzing extensive collections of administrative documents. This approach allowed for the swift identification of both current policies and emerging needs, demonstrating that tasks which traditionally require extensive human effort can now be executed in a fraction of the time and with negligible financial cost. The findings from this study are significant in several respects. First, the application of GPT-4 revealed its capability to generate comprehensive lists of preliminary policies, which were subsequently refined to a more targeted selection relevant to each region. This capability not only streamlined the policy analysis process but also enhanced the precision of policy identification, thereby facilitating more informed decision-making in public administration. Moreover, the study introduced a novel integration of AI tools within the realm of policy analysis, marking a departure from conventional methodologies that rely heavily on manual effort. This innovation has the potential to revolutionize the field by enabling more dynamic and responsive governance models that can anticipate and adapt to emerging societal needs more effectively.

For future works, an important direction involves further refinement of the AI models to improve their ability to contextualize and interpret the nuances of policy documents across different regions and policy areas. Additionally, expanding the application of these models to include real-time analysis of ongoing policy discussions could enhance the ability of policymakers to respond proactively to changing circumstances. From a computational perspective, future work could focus on enhancing the robustness and scalability of the AI algorithms used in policy analysis. This could involve the development of more sophisticated techniques for handling large datasets, improving the interpretability of AI-generated outputs, and ensuring that the models are able to generalize effectively across different contexts. Additionally, integrating these AI tools with other data sources, such as demographic and economic data, could provide a more comprehensive basis for policy formulation and evaluation. In conclusion, this study highlights the transformative potential of AI in the domain of policy analysis and anticipatory governance. By continuing to refine and expand these tools, we can progress towards a more agile and responsive model of governance, better equipped to address the complex challenges of the future.

References

Bolzan, M. (2018). *Domani in famiglia*. Franco Angeli.

Calleo, Y., Pilla, F. (2023). Delphi-based future scenarios: A bibliometric analysis of climate change case studies. *Futures*, **149**, 103143.

Di Zio, S., Calleo, Y., Bolzan, M. (2023). Delphi-based visual scenarios: An innovative use of generative adversarial networks. *Futures*, **154**, 103280.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv*:1909.00512.

Fuerth, L. S. (2009). Foresight and anticipatory governance. *Foresight*, **11**(4), pp. 14-32.

Ha, J. H., Lee, D. H., Choi, B. J. (2023, November). Text mining based GPT method for analyzing research trends. In *International Conference on Intelligent Human Computer Interaction*. Cham: Springer Nature Switzerland, pp. 26-31.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., ... Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv*:2304.10428.

doi.org/10.26398/asaproc.0083

A non-parametric test for comparative evaluations of students' performances

Stefano Bonnini[a], Michela Borghesi[a], Massimiliano Giacalone[b], Gianfranco Piscopo[c]

[a] Department of Economics and Management, University of Ferrara, Ferrara, Italy.
[b] Department of Economics, University of Campania "Luigi Vanvitelli", Capua, Italy.
[c] Department of Mathematics and Applications "Renato Caccioppoli", University of Naples "Federico II", Napoli, Italy.

1. Introduction

Comparing student performance is an effective way of educational evaluation. One crucial tool for evaluating the efficacy of educational practices concerns a statistical analysis applied to student performances before and after a course (Dunlosky et al., 2013). This analysis serves as a means to gauge the impact of instructional strategies on students' learning outcomes and to inform pedagogical decisions. The relevance of such statistical examinations lies in their ability to provide empirical evidence of the effectiveness of specific educational interventions. By comparing students' scores before and after a course, it becomes possible to assess not only the extent of students' academic growth but also the influence of specific instructional techniques and other types of factors on their learning trajectories (Cook and DeMets, 2007). In the literature, some techniques related to data mining or algorithms have been applied to assess the level of education and predict students' performance (Zhang et al., 2021; Tomasevic et al., 2020).

In this work, the approach is based on a non-parametric inference in an experimental framework where the response variable is the student's score variation (score after the course minus score before the course). The goal is to investigate the main and interaction effects of crucial factors such as the type of school (public or private) and the didactic approach (traditional or experimental). The problem is defined in the framework of a regression analysis and the solution is based on the permutation approach. The proposed test, unlike the parametric approaches, does not require the assumption that the distribution of the response follows a specific family of probability laws. Such a test is powerful, especially (but not only) when the typical assumptions of the parametric approaches (such as the normality of data) are not satisfied and the parametric tests are not reliable (Pesarin, 2001). In other words, the probability of correct decision in the test of hypothesis is usually higher than that of the parametric competitors. Furthermore, this method is more flexible and robust concerning the parametric tests and it could be considered in contrast with (but not only) stepwise regression (Harrar and Bathke, 2008). A permutation test on the goodness-of-fit of a multiple regression model is applied. Such a test is conceived as a multiple test on the significance of the single regression coefficients (main effects and interaction effects), and the proposed solution is based on the combination of the p-values of the partial tests. The empirical studies in which this nonparametric methodology has been successfully applied are several and heterogeneous (Bonnini et al., 2023; 2024a; 2024b; Alibrandi et al., 2022).

The rest of the paper is organized as follows. Section 2 presents the statistical problem and the methodological proposed solution. The case study is described in Section 3 and Section 4 contains the overall conclusions.

2. Statistical problem and methodological solution

Let us consider a model with binary variables as predictors, to represent the main effects and the interaction effects of two factors with two or more levels. This situation was first considered by Hollander and Wolfe (1999) because it is typical of experimental designs where subjects are first

divided into homogeneous subgroups (blocks) and then randomly assigned to various treatment levels. Under the null hypothesis, all the treatment effects are null. On the other hand, in the alternative hypothesis, at least one effect is not equal to zero.

With obvious notation, the data can be represented by $y_1, y_2, \dots, y_n$, where n is the total number of observations and y_i is the ith observation of the response, with $i = 1, \dots, n$. If we consider the case of two factors at two levels, i.e. $K = C = 2$, and assume y_i to be a realization of the random variable Y_i, an alternative linear model in the framework of the linear regression analysis is the following:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_{12} D_{i1} D_{i2} + \varepsilon_i,$$

where D_{iv} is a dummy variable that takes the value 0 if the ith observation is associated with the vth factor at level 1, and takes the value 1 if the ith observation is related to the vth factor at level 2, with $i = 1, \dots, n$ and $v = 1,2$. In the regression model, $\varepsilon_1, \dots, \varepsilon_n$ are exchangeable random errors and $\beta_0, \beta_1, \beta_2, \beta_{12}$ are unknown parameters, with β_0 expected value of the response corresponding to the baseline (both the factors at level 1), β_1 and β_2 main effect of factor 1 and factor 2 respectively, and β_{12} interaction effect. We are interested in the following system of hypotheses:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_{12} = 0 \\ H_1: \overline{H_0} \end{cases}.$$

If, instead of testing all the parameters jointly, we were interested in only one of them, under the null hypothesis, exchangeability would hold only between specific data blocks and only synchronized permutations would be allowed (Pesarin and Salmaso, 2010). This is not the case, because, under H_0, all the effects are null and exchangeability holds without constraints or conditions.

The proposed solution consists of the application of permutation tests on the significance of the single coefficients and then a combined permutation test (CPT) to solve the overall goodness-of-fit testing problem. The test statistics of the partial tests on the regression coefficients are the absolute values of suitable estimators of the parameters (such as the least square estimators): $T_1 = |\hat{\beta}_1|, T_2 = |\hat{\beta}_2|$ and $T_{12} = |\hat{\beta}_{12}|$, respectively. The null distribution of the test statistics is obtained by permuting the rows of the design matrix, keeping the vector of observed values of the response fixed, and recalculating for each permutation the estimates of the coefficients and, consequently, the values of the test statistics.

This non-parametric method is very useful for solving complex problems and the main advantage concerning other parametric methods is that the multivariate distribution of the test statistic does not need to be known, and in particular, the dependence structure between variables does not need to be explicitly modelled or specified (Bonnini et al., 2014). Permutation tests are distribution-free, hence they are flexible and robust concerning the departure from normality (Pesarin and Salmaso, 2010).

The dependence of the partial tests is implicitly taken into account through the permutation of the rows of the design matrix. By assuming, without loss of generality, that the overall hypothesis is rejected for large values of the test statistic, the following combination rule can be used to compute the combined test statistic for the overall problem:

$$T_{comb} = \max(1 - \lambda_1, 1 - \lambda_2, 1 - \lambda_{12})$$

where λ_1, $\lambda_{2,}$ and λ_{12} are the p-values of the partial tests. Since such a permutation ANOVA for the linear regression models is defined as a multiple test, in the case of a rejection of the null hypothesis in favor of the alternative and, in order to attribute the overall significance to specific

partial tests (i.e., to specific coefficient estimates), the control of the family wise error (FWE) is necessary (Westfall and Young, 1992; Westfall and Young, 1989). In other words, to avoid the inflation of the type I error of the overall test, we must adjust the partial p-values. A suitable method is the one based on the Bonferroni-Holm rule (Giacalone et al., 2018). The regression model under consideration for our case study is the following:

$$y_i = \beta_0 + \beta_1 \cdot type_school_i + \beta_2 \cdot teaching_method_i + \beta_{12} \cdot interaction_i + \varepsilon_i$$

where y_i is the score difference for the ith student, $type_school_i$ is the dummy variable that takes the value 1 if the ith student is in a public school and 0 otherwise, $teaching_method_i$ is a dummy variable that takes the value 1 if the ith student attended a course with an experimental teaching approach and 0 otherwise, and $interaction_i$ is the product between $type_school_i$ and $teaching_method_i$, hence it takes the value 1 if both the dummy variables take the value 1, in other words if the level of both factors changes with respect to the baseline (from private school and standard teaching approach to public school and experimental teaching approach).

3. Simulation study

To study the power behaviour of the proposed nonparametric solution in comparison with the parametric F-test for ANOVA, a Monte Carlo simulation study was carried out. The application of the test, the data random generations, and the comparative performance assessments were carried out through new R scripts created by the authors.

In the simulation study, 1000 datasets were simulated under two different settings. In one setting, all the regression coefficients are not equal to zero; in the other setting, only one regression coefficient (main effect of a factor) and the intercept are different from zero. The latter setting simulates a scenario similar to the case study described in the next section. We also took inspiration from the case study to simulate the predictors' data. Such data were obtained as a simple random sample of the rows of the design matrix X. To prove the superiority of the proposed method under asymmetric non-normal errors (with null mean and a variance similar to the one of the case study), n i.i.d. random errors were generated from a chi-square distribution with 3 degrees of freedom. The values of the dependent variable were obtained by summing the simulated deterministic part of the model and the standardized simulated errors multiplied by $\sqrt{11}$.

In Figure 1 and Figure 2 the rejection rates of the parametric ANOVA of the regression model based on the F test and the proposed non-parametric method based on the CPT approach are reported as a function of the sample size, under the two settings. It is evident that in both cases, with large sample sizes, the two methods are equivalent, while for small sample sizes, the non-parametric solution is preferable.

Figure 1. Rejection rates of the two tests under H_1 as function of the sample size, with $\beta_0 = 10, \beta_1 = \beta_2 = \beta_{12} = 2, \alpha = 0.05$.

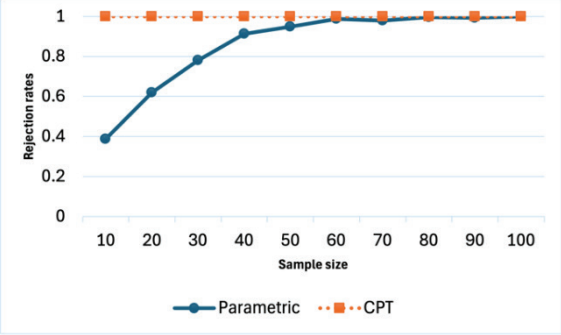
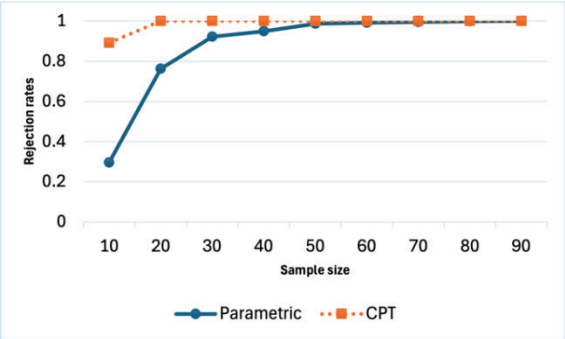


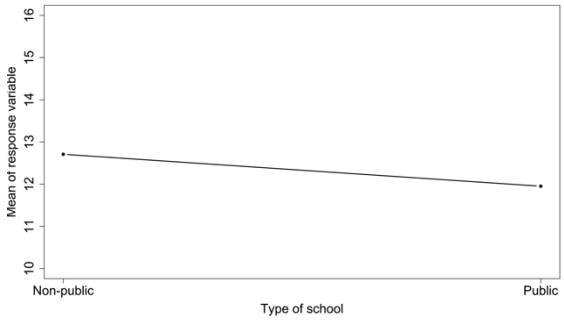
Figure 2. Rejection rates of the two tests under H_1 as a function of the sample size, with $\beta_0 = 10, \beta_1 = \beta_{12} = 0, \beta_2 = 5, \alpha = 0.05$.

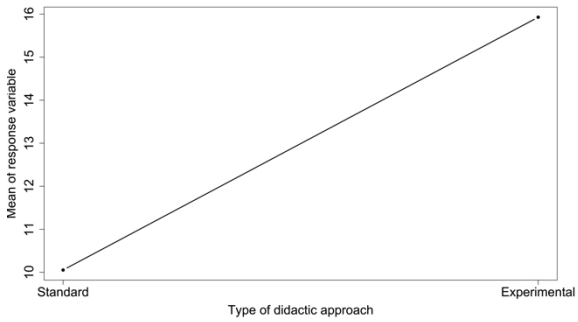


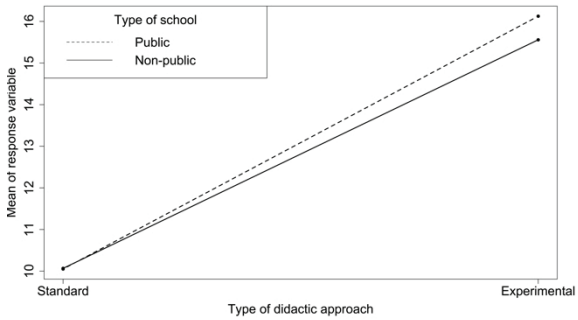
4. Case study

The case study discussed in this work relates to a dataset about students' performance. The data source is Kaggle (https://www.kaggle.com/datasets/kwadwoofosu/predict-test-scores-of-students), a data science competition platform and online community for data scientists and machine learning practitioners. This dataset is part of SPSS v23 package and the aim concerns an American district school board wanting to know if an experimental teaching method effectively improves math scores. They select at random schools within the district, then some classrooms within each school. The classrooms were randomly assigned to the experimental or control group, and the classroom teachers were asked to use the new or standard method to teach mathematics. All classroom students were tested at the beginning and end of the school year. The random assignment of "treatments" to students implies exchangeability. Hence, the 2133 students involved in this study took an exam both before and after attending a course in Mathematics. For each student, the pre-score and the post-score, i.e. the score obtained in the exam before and after the course, were observed, and the difference between the two values was used to measure the improvement of the learning skills. Such a difference represents the response variable. The other variables of the model are defined in section 2. In Figure 3, the main effect plots and the interaction plot are represented. From a descriptive point of view, a variation in the means in both the main effect plots is shown, even much more evident in the case of the didactic approach, while in the interaction plot, the two segments are not parallel even if the difference in slopes is not very evident.

Figure 3. Main effect plots and interaction plot.







The goal of the analysis is to test the system of hypotheses defined above. To this aim, the nonparametric test presented was applied to such data. The significance level was set equal to 0.01. The global p-value of the test is equal to 0.0003, thus it is less than the significance level, which leads to the rejection of the null hypothesis of no effect. The partial p-values related to the tests on the significance of the coefficient estimates adjusted with the Bonferroni-Holm correction are shown in Table 1. According to the adjusted p-values, the significance of the overall test can only be attributed to one of the regression coefficients which corresponds to a main effect. In particular, it appears that only the teaching method, when it changes from standard to experimental, can positively influence the test score.

Table 1. Estimates of the coefficients, partial p-values on the significance of the estimates, adjusted with the Bonferroni–Holm method (significance in bold).

	Coefficients	**Adjusted p-values**
Intercept	10.066	
Type of school	-0.017	0.9521
Teaching method	5.492	**0.0003**
Interaction	0.584	0.3510

5. Concluding remarks

This work aimed to evaluate the efficacy of educational practices by analyzing student performance data before and after administering a given course. The final goal is to gauge the impact of instructional strategies on students' learning outcomes to make suitable pedagogical decisions. The usefulness of such educational experiments lies in the opportunity to assess the effectiveness of specific educational interventions. A combined permutation test allowed us to test the significance of the main and interaction effects of factors such as the type of school and the teaching method. The scientific contributions of the work consist in the application of this innovative and performant method to an interesting case study of educational statistics and the empirical result that enriches the literature on the topic. Specifically, we had empirical evidence of the importance of the teaching method whereas the type of school (public or private) has no effect.

A future extension of this work could explore the method's performance on datasets with more complex dependency structures or alternative parametric models.

References

Alibrandi, A., Giacalone, M., Zirilli, A. (2022). Psychological stress in nurses assisting amyotrophic lateral Aclerosis patients: A statistical analysis based on non-parametric combination test. *Mediterranean Journal of Clinical Psychology*, **10**(2).

Bonnini, S., Borghesi, M., Giacalone, M. (2023), Advances on multisample permutation tests for "V-shaped" and "U-shaped" alternatives with application to circular economy. *Annals of Operations Research*, pp. 1-16.

Bonnini, S., Borghesi, M., Giacalone, M. (2024a) Semi-parametric approach for modeling overdispersed count data with application to industry 4.0. *Socio-Economic Planning Sciences*, 101976.

Bonnini, S., Borghesi, M., Giacalone, M. (2024b) Simultaneous marginal homogeneity versus directional alternatives for multivariate binary data with application to circular economy assessments. *Applied Stochastic Models in Business and Industry*; **40**(2): pp. 389–407.

Bonnini, S., Corain, L., Marozzi, M., Salmaso, L. (2014). *Nonparametric Hypothesis Testing, Rank and Permutation Methods with Applications in R*; Wiley: Hoboken, NY, USA.

Cook, T. D., DeMets, D. L. (2007). *Introduction to Statistical Methods for Clinical Trials*. Chapman and Hall/CRC.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, **14**(1), pp. 4-58.

Giacalone, M., Agata, Z., Cozzucoli, P. C., Alibrandi, A. (2018). Bonferroni–Holm and permutation tests to compare health data: Methodological and applicative issues. *BMC Medical Research Methodology*, **18**(1), pp. 1–9.

Harrar, S.W., Bathke, A.C. (2008). A non-parametric version of the Bartlett-Nanda-Pillai multivariate test. Asymptotics, approximations, and applications. *American Journal of Mathematical and Management Sciences*, **28**, pp. 309–335.

Hollander, M., Wolfe, D.A. (1999) *Nonparametric Statistical Methods*. John Wiley & Sons, Ltd.

Pesarin, F. (2001). *Nonparametric Combination Methodology. In Multivariate Permutation Tests with Applications in Biostatistics*, 2nd ed.; Wiley: Chichester, UK.

Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*; Wiley: Hoboken, NY, USA.

Tomasevic, N., Gvozdenovic, N., Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, **143**, 103676.

Westfall, P.H., Young, S.S. (1992). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*; Wiley-Interscience: New York, NY, USA.

Westfall, P.H., Young, S.S. (1989). P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, **84**, pp. 780–786.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. *Frontiers in Psychology*, **12**, 698490.

doi.org/10.26398/asaproc.0084

An empirical analysis of various methods for making inferences with non-probability samples

Lisa Braito, Emilia Rocco
Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy.

1. Introduction

Non-probability sampling has become increasingly prevalent across various fields, especially in social studies and policy research, due to rapid access to numerous low-cost, non-probability data sources like web surveys and social media. However, unlike probability sampling, the chance of being selected is unknown in non-probability sampling, and controlled random selection is absent. This makes traditional design-based methods of inference, commonly used in probability surveys, inadequate and raises concerns about potential bias due to self-selection.

To ensure valid inferences with non-probability samples, various methods have been developed over time. A review of them is given in (Wu, 2022). Some of these methods adapt techniques used in probability sampling to address non-response issues. Furthermore, some of these methods align with those used in the causal inference context, where data from randomized controlled trials and observational studies are combined to enhance the generalizability of the former by leveraging the representativeness of the latter. Despite their differences, all the proposed methods rely on the use of auxiliary information available at the population level or, in most cases, from a probability sample drawn from the same study population. They can be categorized based on whether auxiliary variables are used to create predictive models for the response variable, models for survey participation propensity, or both, using techniques such as doubly robust estimation or calibration. For both the response variable and the propensity score a range of parametric and non-parametric models, explicit or implicit, are possible. Our goal is to empirically compare some of these methods in contexts involving multiple auxiliary variables with diverse characteristics and relationships, both among themselves and with the study variable, while considering various self-selection processes. Indeed, unlike probability sampling, non-probability sampling does not have a single definition. Generally, any subset of the population that does not meet the criteria for probability sampling is considered a non-probability sample. Consequently, self-selection processes can vary widely.

The paper is structured as follows: Section 2 introduces the notation and briefly reviews the methods for making inferences with non-probability samples, which are empirically compared in a large simulation study in Section 3. Finally, Section 4 concludes with a discussion.

2. Notation and methods

Let $U = \{1, \dots, N\}$ be a finite population of N units and $(\boldsymbol{x_i}, y_i)$ be the measurements of the p auxiliary variables $\boldsymbol{X}$ and the study variable Y associated with unit $(i = 1, \dots, N)$. We are interested in estimating $\mu_y = \sum_{i=1}^{N} y_i / N$. The sources of data are two: a non-probability sample s_A with information on both the study variable Y and the auxiliary variables $\boldsymbol{X}$ and a probability sample s_B with observations only on the auxiliary variables $\boldsymbol{X}$. While for s_B the inclusion probabilities $\pi_i^B = P(i \in s_B)$ are known, for s_A the inclusion probabilities π_i^A cannot be calculated because the sampling mechanism is unknown. Consequently, μ_y cannot be estimated through the Horvitz-Thompson estimator and the sample mean (i.e. the naive estimator) applied to s_A without adjusting

for the sampling process are usually subject to selection biases. The biased nature of non-probability samples cannot be corrected by using the sample itself. Additional auxiliary information about the target population is required. When population-level data is unavailable, but a probability sample (s_B) from the same finite population exists, this information can be derived from it. However, all methods for making inferences with a non-probability sample by leveraging auxiliary information shared with a probability sample rely on specific assumptions. The three key assumptions common to all these methods are:

I. The sample inclusion indicator variable of the non-probability sample, $R_i = I(i \in s_A)$, and the outcome variable y_i are independent given the set of auxiliary variables, that is $(R_i \perp y_i)|\boldsymbol{x}_i$. This assumption is analogous to the missing at random (MAR) assumption used in missing data analysis.

II. Every unit in the target population has a non-zero probability of being included in the non-probability sample; formally $\pi_i^A > 0, (i = 1, \dots, N)$.

III. The participation of units in the non-probability sample (represented by indicator variables $R_1, \dots, R_N$) is independent of one another, given the set of auxiliary variables.

Other assumptions are specific for each method or category of methods. Indeed, as just noted in Section 1 the different methods can be classified into three categories.

In the first category the probability of a unit being included in the non-probability sample, which is referred to as the propensity score, is modelled and estimated for all units in the non-probability sample. Under assumption (I), the propensity scores $\pi_i^A = P(R_i = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ are a function of the auxiliary variables $\boldsymbol{X}$ but their functional form may vary. In this paper, only the parametrical logit model $\pi_i^A = exp(\boldsymbol{x}_i'\beta)\big(1 + exp(\boldsymbol{x}_i'\boldsymbol{\beta})\big)^{-1}$ is considered. As in Chen et al (2020), the parameters' vector $\boldsymbol{\beta}$ is estimated using the maximum pseudo-likelihood method. This method, which incorporates the design weights, enables the production of a consistent estimator for the parameter vector $\boldsymbol{\beta}$ in the context of a complex survey design. The vector $\widehat{\boldsymbol{\beta}}$ is then used to calculate $\hat{\pi}_i^A$, $i \in s_A$. Finally, the inverse probability weighted estimator $\hat{\mu}_{IPWy}$ of μ_y is constructed as

$$\hat{\mu}_{IPWy} = \frac{\sum_{i \in s_A} y_i / \hat{\pi}_i^A}{\sum_{i \in s_A} 1/\hat{\pi}_i^A}$$

Other specifications of the propensity score models are possible.

The most crucial assumption of methods in the second category is the definition of a prediction model $f(y|x)$ which can be estimated by using the observed $(\boldsymbol{x}_i, y_i)$ from the non-probability sample s_A. The estimated prediction model is then used to create mass-imputed values $\widehat{y}_i$ using observed $\boldsymbol{x}_i$ for all the units in the probability sample s_B. The mass imputation estimator of μ_y is finally constructed as

$$\hat{\mu}_{MIy} = \frac{\sum_{i \in s_B} \widehat{y}_i / \pi_i^B}{\sum_{i \in s_B} 1/\pi_i^B}$$

In what follows we consider two possible choices for $f(y|x)$: a generalized regression model and a generalized additive model (Hastie and Tibshirani, 1990). In the following, the two corresponding mean estimators are denoted as $\hat{\mu}_{MI(P)y}$ and $\hat{\mu}_{MI(NP)y}$ respectively. Other possible specifications of $f(y|x)$ not considered in this paper include the nearest neighbour imputation and the use of machine learning methods such as random forests.

The latter category includes the methods in which both the propensity score model and the outcome regression model are considered. These methods are termed double robust (DR) since the corresponding mean estimator is consistent if either the propensity score model or the outcome regression model is correctly specified. In particular, the two models can be separately defined as above described. Then they are combined in the following mean estimator expression

$$\hat{\mu}_{DRy} = \widehat{N}_A^{-1} \sum_{i \in s_a} \frac{1}{\hat{\pi}_i^A} \{y_i - \widehat{y}_i\} + \widehat{N}_B^{-1} \sum_{j \in s_B} \frac{\widehat{y}_j}{\pi_j^B} \quad (1)$$

where $\widehat{N}_A = \sum_{i \in s_A} \left(\hat{\pi}_i^A\right)^{-1}$ and $\widehat{N}_B = \sum_{j \in s_B} \left(\pi_j^B\right)^{-1}$. We have considered in this study one specification for the propensity score model and two different specifications for the outcome model, parametric and nonparametric as defined above. In the following, the two corresponding mean estimators are denoted as $\hat{\mu}_{DR(P)y}$ and $\hat{\mu}_{DR(NP)y}$ respectively.

Even though calibration is often considered a stand-alone method, it is actually a specific type of DR method because it implicitly assumes both a propensity score model and an outcome regression model. The calibration strategy assigns a weight ω_{Ai} to each unit $i \in s_A$ so that

$$\sum_{i \in s_A} \omega_{Ai}\, x_{ik} = \sum_{j \in s_B} d_{Bj}\, x_{jk} \quad k = 1, \ldots, p \quad (2)$$

where $d_{Bj} = 1/\pi_j^B$. These weights are obtained by solving the following optimization problem:

$$min_{\omega_{Ai}: i \in s_A} \left(\sum_{i \in s_A} \omega_{Ai}\, log \omega_{Ai} \right)$$

subject to $\omega_{Ai} \geq 0$, for all $i \in s_A$; $\sum_{i \in s_A} \omega_{Ai} = N$, and the balancing constraint (2). Lagrange multiplier method can be used to solve this optimisation problem. The estimated weights are then used to calculate the calibration mean estimator

$$\hat{\mu}_{CALy} = \frac{1}{N} \sum_{i \in s_A} \omega_{Ai}\, y_i$$

No model has been explicitly defined, however, the justification for $\hat{\mu}_{CALy}$ relies on the linearity of the outcome model and a log-linear model for the sampling propensity score. Other objective functions of the weights correspond to different models for the propensity scores. Different models for the outcome variable can be accomodated by considering more complex specifications of the balancing constraint (2) but this potentially complicates the solution of the optimization problem. More recently, Lee et al. (2023), to obtain double robustness of calibrated estimator in situations in which the outcome model is not linear propose to combine the calibration approach just described with a prediction model for the outcome variable defining a mean estimator with the same structure of the double robust one in (1). We call this estimator augmented calibration estimator (ACAL), and its expression is:

$$\hat{\mu}_{ACALy} = \frac{\sum_{i \in s_A} \omega_{Ai} \{y_i - \widehat{y}_i\}}{\sum_{i \in s_A} \omega_{Ai}} + \widehat{N}_B^{-1} \sum_{j \in s_B} \frac{\widehat{y}_j}{\pi_j^B}$$

In this case as well, we considered both the parametric and non-parametric models for the outcome variable, with the corresponding estimators denoted as $\hat{\mu}_{ACAL(P)y}$ and $\hat{\mu}_{ACAL(NP)y}$ respectively.

3. Simulation study description and results

This section describes the simulation study, which aims to compare how different inference approaches for non-probability samples perform in various situations, and consequently considers different types of auxiliary and response variables. A unique population of $N = 20000$ is considered. The auxiliary variables are five and are generated as follows: $\forall i = 1, \ldots, N$ $x_{1i} = z_{1i}$ with $z_{1i} \sim \text{Bernoulli}(0.7)$, $x_{2i} = (1, 2, 3)^T \boldsymbol{z}_{2i}$ with $\boldsymbol{z}_{2i} \sim \text{Multinomial}_3(n = 1;\ 0.2, 0.5, 0.3)$, $x_{3i} = z_{3i} + 0.3\big((10(z_{1i} + 0.15)z_{2i})/10\big)$ with $z_{3i} \sim \text{Uniform}(0, 2)$, $x_{4i} = z_{4i} + 0.1z_{3i} + 0.1z_{1i}$ with $z_{4i} \sim \text{Exponential}(1)$ and $x_{5i} = z_{5i}$ with $z_{5i} \sim \text{Poisson}(2)$.

We assume to be interested in 7 outcome variables generated by using the following continuous and non-continuous, linear and non-linear outcome models:

a. $y_{1i} = \boldsymbol{\beta}_0^T \boldsymbol{X}_i + \sigma\epsilon_i$, with $\epsilon_i \sim \mathrm{N}(0,1)$, with $\boldsymbol{\beta}_0 = (2,1,1,1,1,1)$ and σ chosen so that the correlation coefficient between y and the linear predictor $\boldsymbol{\beta}_0^T \boldsymbol{X}_i$ is $\rho = 0.8$
b. $y_{2i} = \boldsymbol{\beta}_0^T \boldsymbol{X}_i + \sigma\epsilon_i$, with $\epsilon_i \sim \mathrm{N}(0,1)$, with $\boldsymbol{\beta}_0 = (2,1,1,1,1,1)$ and σ chosen so that the correlation coefficient between y and the linear predictor $\boldsymbol{\beta}_0^T \boldsymbol{X}_i$ is $\rho = 0.3$
c. $y_{3i} \sim \mathrm{Ber}\{\pi_{y_3}(\boldsymbol{X})\}$ with $\mathrm{logit}\{\pi_{y_3}(\boldsymbol{X})\} = \boldsymbol{\beta}_0^T \boldsymbol{X}_i$ with $\boldsymbol{\beta}_0 = (-2,1,1,-1,1,1)$
d. $y_{4i} \sim \mathrm{Pois}\{\lambda_{y_4}(\boldsymbol{X}) + 4\}$ with $\lambda_{y_4}(\boldsymbol{X}) = \boldsymbol{\beta}_0^T \boldsymbol{X}_i$ with $\boldsymbol{\beta}_0 = (2,1,1,1,1,1)$. This Poisson distribution is symmetric ($skewness = 0.4$, $kurtosis = 0.2$).
e. $y_{5i} \sim \mathrm{Pois}\{\lambda_{y_5}(\boldsymbol{X}) + 4\}$ with $\lambda_{y_5}(\boldsymbol{X}) = \boldsymbol{\beta}_0^T \boldsymbol{X}_i$ with $\boldsymbol{\beta}_0 = (-2,1,1,-1,1,1)$. This Poisson distribution is asymmetric ($skewness = 0.8$, $kurtosis = 0.8$).
f. $y_{6i} = 5 + \exp\{3\sin(\boldsymbol{\beta}^T \boldsymbol{X}_i)\} + \epsilon_i$, with $\epsilon_i \sim \mathrm{N}(0,1)$ and with $\boldsymbol{\beta} = (1,1,-1,0.9,0.9)$ and
g. $y_{7i} \sim \mathrm{Ber}\{\pi_{y_7}(\boldsymbol{X})\}$ with $\mathrm{logit}\{\pi_{y_7}(\boldsymbol{X})\} = \boldsymbol{\beta}_0^T \widetilde{\boldsymbol{X}_i}$ with $\boldsymbol{\beta}_0 = (-2,1,1,1,-1,1)$ and $\widetilde{\boldsymbol{X}_i} = (x_{1i}, x_{2i}, x_{3i}^2, x_{4i}, x_{5i})$

Regarding the self-selection process of the non-probability sample, we consider two models:

s1 $\mathrm{logit}(\pi_{1i}^A) = \boldsymbol{\alpha}_0^T \boldsymbol{X}_i$, where $\boldsymbol{\alpha}_0 = (-4.44, 0.1, 0.1, 0.1, 0.2, 0.1)$ where the constant is chosen such that the $\sum_{i=1}^N \pi_{1i}^A = n_A$ with the given target sample n_A
s2 given $\eta_i = \boldsymbol{\alpha}^T \boldsymbol{X}_i$ with $\boldsymbol{\alpha} = (0.1, 0.1, 0.1, 0.2, 0.1)$, $\pi_{2i}^A = 0.01$ if $\eta_i < \mathrm{median}(\eta_i)$ or $\pi_{2i}^A = 0.04$ if $\eta_i \geq \mathrm{median}(\eta_i)$

Given the specified setting, we conduct $B = 5000$ replications of the probability and non-probability sample selection processes and the subsequent inference procedures. The non-probability sample s_A is selected from the population by fixing an expected sample size of 500 and using the Poisson sampling method, for each of the two self-selection processes. In contrast, the probability sample s_B is a simple random sample (SRS) of size $n_B = 1000$.

To estimate the mean of the 7 outcome variables assumed to be known only for the units included in the non-probability sample, we consider four scenarios for inference, corresponding to the correct specification of at least one of the two models (the model for the propensity to be in the non-probability sample and the model for the outcome) as well as the scenario where neither model is correctly specified.

We evaluate the performance of various point estimators for μ_y versus the naive estimator $\frac{1}{n_A}\sum_{i=1}^{n_A} y_i$ (sample mean of s_A). The performance is evaluated through the relative bias in percentage (RB%) and the mean squared error (MSE) computed on the B simulation runs:

$$\mathrm{RB\%} = \frac{1}{B}\sum_{b=1}^{B} \frac{\hat{\mu}^b - \mu_y}{\mu_y}, \ \mathrm{MSE} = \frac{1}{B}\sum_{b=1}^{B} \left(\hat{\mu}^b - \mu_y\right)^2$$

where $\hat{\mu}^b$ is the estimator computed from the bth selection of the two samples. Simulation results for the four scenarios for all the estimators are reported in Table 1 and Table 2.

As shown in Table 1, when both the outcome and selection models are well specified, all the estimators are robust and comparable. In particular, for the first two continuous response variables, the inherent double robustness of the calibration estimator is evident. Its RB% and MSE match exactly with those of the double robust estimator. An exception is observed in the TT results for the variable Y_5, where the mass imputation parametric method performs worse than the others. This is likely due to the asymmetry of the variable, which may imply that the estimation of the outcome model parameters is more sensitive to the self-selection of the non-probability sample. In all cases, for the Poisson responses, the semiparametric approach considered seems to have limitations. Different non-parametric models need to be explored.

Looking at the last two columns of Table 1, it is evident that, as expected, $\hat{\mu}_{MI(P)y}$ performs the worst. The semiparametric approach ($\hat{\mu}_{MI(NP)y}$) generally improves the RB%s. However, this improvement is more noticeable for variable Y_7, which has a polynomial functional form,

compared to variable Y_6, which has a more complex non-linear functional form. Different non-parametric approaches may offer better performance in such cases.

	True Outcome Model & True selection model (TT)					False Outcome Model & True selection model (FT)	
Estimator	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
$\hat{\mu}_{NAIVE}$	6.85868	5.73564	7.6265	4.46201	8.28594	8.28594	6.4329
	(0.33244)	(0.16238)	(0.00295)	(0.33255)	(0.19733)	(0.49119)	(0.00148)
$\hat{\mu}_{IPWy}$	-0.09119	-0.1185	-0.0354	-0.05571	-0.1166	-0.1166	-0.07559
	(0.01177)	(0.03482)	(0.00048)	(0.03225)	(0.01411)	(0.10849)	(0.00045)
$\hat{\mu}_{CALy}$	-0.0349	-0.07111	0.05343	-0.01853	-0.04681	-0.04681	0.01164
	(0.00975)	(0.03386)	(0.00043)	(0.03001)	(0.01276)	(0.10127)	(0.00041)
$\hat{\mu}_{MI(P)y}$	-0.03426	-0.06912	0.06137	0.0958	0.3059	0.3059	0.22714
	(0.00972)	(0.03374)	(0.00042)	(0.02980)	(0.01296)	(0.10402)	(0.00042)
$\hat{\mu}_{DR(P)y}$	-0.03508	-0.07166	0.02771	-0.02023	-0.03817	-0.03817	-0.01448
	(0.00975)	(0.03384)	(0.00042)	(0.03025)	(0.01327)	(0.10243)	(0.00041)
$\hat{\mu}_{MI(NP)y}$	-0.03255	-0.06364	0.04312	0.11989	0.16975	0.16975	0.03216
	(0.00982)	(0.03429)	(0.00043)	(0.13752)	(0.01306)	(0.09915)	(0.00041)
$\hat{\mu}_{DR(NP)y}$	-0.03267	-0.06412	0.02048	0.02853	-0.01167	-0.01167	0.00076
	(0.00983)	(0.03437)	(0.00043)	(0.13785)	(0.01333)	(0.10069)	(0.00041)
$\hat{\mu}_{ACAL(P)y}$	-0.0349	-0.07111	0.02966	-0.01533	-0.02292	-0.02292	-0.00498
	(0.00975)	(0.03386)	(0.00042)	(0.03017)	(0.01315)	(0.10127)	(0.00041)
$\hat{\mu}_{ACAL(NP)y}$	-0.03259	-0.06385	0.02173	0.03218	-0.00407	-0.00407	-0.00813
	(0.00984)	(0.03439)	(0.00043)	(0.13779)	(0.01327)	(0.09739)	(0.00041)

Table 1: Simulated RB% and MSE of estimators of μ_y - TT and FT scenarios

	True Outcome Model & False selection model (TF)					False Outcome Model & False selection model (FF)	
Estimator	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
$\hat{\mu}_{NAIVE}$	11.54167	9.01848	14.61845	7.5545	12.94779	-12.87091	9.47334
	(0.91064)	(0.34798)	(0.00972)	(0.88766)	(0.45647)	(2.24937)	(0.00261)
$\hat{\mu}_{IPWy}$	2.51803	2.06709	5.29971	1.68261	3.16718	-5.72281	4.05522
	(0.05704)	(0.05357)	(0.00177)	(0.07844)	(0.04202)	(0.54398)	(0.00091)
$\hat{\mu}_{IPWy}$	0.00202	-0.01891	1.74976	0.02295	-0.12543	-2.6782	0.52798
	(0.01075)	(0.03881)	(0.00065)	(0.03359)	(0.01417)	(0.22353)	(0.00051)
$\hat{\mu}_{MI(P)y}$	0.0095	0.00448	0.05475	0.62409	1.62171	-4.85367	1.28564
	(0.01043)	(0.03706)	(0.00053)	(0.03710)	(0.02023)	(0.41591)	(0.00052)
$\hat{\mu}_{DR(P)y}$	0.00645	-0.00503	-0.01067	0.37207	1.07725	-3.59676	0.53279
	(0.01055)	(0.03774)	(0.00053)	(0.03465)	(0.01704)	(0.28873)	(0.00049)
$\hat{\mu}_{MI(NP)y}$	-0.00601	-0.04278	-0.00523	1.11734	1.09908	-1.94996	0.24766
	(0.01070)	(0.03843)	(0.00054)	(8.34234)	(0.01730)	(0.42500)	(0.00050)
$\hat{\mu}_{DR(NP)y}$	-0.00417	-0.03702	-0.05501	0.90568	0.77505	-1.55365	0.08075
	(0.01079)	(0.03899)	(0.00055)	(8.33868)	(0.01619)	(0.41366)	(0.00050)
$\hat{\mu}_{ACAL(P)y}$	0.00202	-0.01891	-0.06287	0.20879	0.7	-2.6782	0.12516
	(0.01075)	(0.03881)	(0.00054)	(0.03412)	(0.01561)	(0.22353)	(0.00049)
$\hat{\mu}_{ACAL(NP)y}$	-0.00484	-0.03915	-0.09781	0.76919	0.52918	-1.18822	-0.01255
	(0.01095)	(0.03987)	(0.00055)	(8.33618)	(0.01565)	(0.40746)	(0.00050)

Table 2: Simulated RB% and MSE of estimators of μ_y - TF and FF scenarios

As shown in Table 2, as expected, the $\hat{\mu}_{IPWy}$ estimator performs the worst. Notably, there is an improvement in the $\hat{\mu}_{ACAL(P)y}$ for the binary response variable Y_3. The calibration estimator is inherently robust for a continuous outcome model, so adjusting it for a binary outcome model shows a significant improvement in the RB%. Unfortunately, this improvement

is not observed for the other non-continuous variables, Y_4 and Y_5. As noted in previous results, mass imputation does not perform well for these Poisson variables. However, Y_4 approximates a normal distribution, allowing the calibration estimator to account for it more effectively. This result requires further investigation. As noted in the literature, both the DR estimator and calibration perform well because they rely on the correctness of just one of the two models. The mass imputation estimator performs well when the outcome model is correct.

Looking at the last two columns of Table 2, the results for Y_7 confirm the previous findings from Table 1, specifically the ability of the semiparametric approach to correctly estimate the outcome model. Y_6 represents our worst-case scenario due to its nonlinear response and the complexity of the selection model. Interestingly, in this case, the best-performing estimator is the augmented calibration estimator, particularly the semiparametric versions.

4. Final remarks

In this study, we compared various methods for making inferences with non-probability samples to assess their robustness across different types of outcome variables and to identify guidelines for selecting the most appropriate method in different situations. All methods rely on the use of auxiliary variables and assume a model either for the relationship between the outcome and the auxiliary variables or for the relationship between the propensity to be included in the non-probability sample and the auxiliary variables. Our analysis is focused on assessing robustness concerning model specifications.

The results show that when the models are correctly specified, all methods perform well and are almost equivalent. When the outcome variable is asymmetric (variable y_5), even if the parametric model for the outcome is correctly specified or a more flexible non-parametric model is used, the mass-imputation approach may be less robust, as the estimates of the model parameters can be affected by the self-selection inherent in the non-probability sample data.

To conclude, the most interesting result is the fact that the augmented semiparametric calibration gives the best performance. Indeed, where everything works it is comparable to the other results, but in case of misspecification of one of the models and especially also in misspecification of all the models, it improves greatly the relative bias. As mentioned previously, this is especially noticeable for the binary response variable, for the Poisson outcome these results are not clear, and we need to further investigate them. Although further investigation is needed, the strong performance of the augmented semiparametric calibration, combined with the general property of calibration to ensure the coherence of estimates across different outcomes, makes this a very promising approach.

Finally, more future theoretical work has to be done on the use of nonparametric models for the specification of the propensity score.

References

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**(532), pp. 2011-2021.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. Wiley Online Library.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., Cai, J. (2023). Improving trial generalizability using observational studies. *Biometrics*, **79**, pp. 1213-1225.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, **48**(2), pp. 283-311.

doi.org/10.26398/asaproc.0085

A statistical approach for the predictive analysis of insolvency in Italy

Eugenio Brentari, Gloria Ermici, Andrea Zanetti
Department of Economics and Management, University of Brescia, Brescia, Italy

1. Introduction

The Italian bankruptcy law has always provided tools to act on companies that were already in a state of crisis, instead of creating a preventive alert mechanism. As a result, the bankruptcy law reform and the crisis code (Legislative Decree 14/2019) were introduced.

This problem has been studied in detail by various authors to create ratings or predictive tools to anticipate the business crisis and the analysis of business continuity and especially the ongoing concern (Chye and Kee, 2004). The going concern principle is "the accountant's firm model" (Sterling, 1968) and is the basis for financial statements. According to the international standard on auditing ISA 570[1], it assumes that the entity is a going concern and will continue its operations for the foreseeable future.

One of the consequences of the lack of ongoing concern is the company's economic and financial disequilibrium which leads the company to bankruptcy. For example, according to ISTAT (2023), the overall number of bankruptcies in Italy in the third quarter of 2023 increased by 5.4% compared to the previous quarter and by 11.4% compared to the third quarter of 2022.

In this work the Altman's Z score model (Altman, 1968, 1993 and 2002; Altman and Hotchkiss, 1995) is proposed to analyse the Italian perspective (Alberici, 1975; Altman, Falini and Danovi, 2013). The new code has delegated to the *Consiglio Nazionale dei Dottori Commercialisti ed Esperti Contabili* (CNDCEC) the task to develop warning indicators able to promptly detect and prevent the state of insolvency of companies in compliance with business continuity.

The CNDCEC determines five more indicators. If all five indicators are out of range, the company will have to communicate the state of crisis to the Chamber of Commerce. The application of the new crisis code, introduced in 2019 and subsequently amended, has seen the extension of the administrator's responsibility, implying the need to adopt adequate organizational, administrative and accounting structures that can detect insolvency situations and to adopt corrective elements. In July 2022, the new crisis code was modified, introducing different indicators, focused on cash and payables. Anyway, the previous indicators remain a tool for companies to check their organizational, administrative and accounting structures.

In the earlier study (Brentari, Ermici, Zanetti, 2024), through a dataset with about 20.000 SMEs with headquarters in Brescia Province, Italy, our previous research tried to verify if these indicators are the right tools to predict the companies' failure. None of the companies in the sample failed in 2020, 2021 and 2022: this point was not enough to declare that indicators are not the right tools to predict the crisis state for Italian SMEs.

In the research, we hypothesized that: 1. The technical and legal times for the bankruptcy declaration by the court of law, especially in 2020 with the suspension introduced by the law decree 23/2020 in the COVID-19 period are beyond the search times; 2. The lack of ongoing concern emerges after the three years analysed, e.g. in 2023 or 2024; and 3. The non-compliance by companies with the reporting obligation and absence of calculation of the indicators.

This research develops the dataset including new and more representative data. This research analyses the companies' behaviour in the subsequent years, and it poses the necessity to question if

[1] International Standard on Auditing 570 (2016), *Going Concern*. https://www.ifac.org/_flysystem/azure-private/publications/files/ISA-570-(Revised).pdf

the macro results obtained from the previous studies have changed. Furthermore, this research analyses Altman's Z score model, together with the SME's version (Altman, Falini and Danovi, 2013), about the companies in the sample and compares these two predictive models.

Hence, this paper aims to verify the following hypothesis: *Are the indicators determined by the Italian law in 2019 the right tools to predict the companies' failure? Is there a connection between the results obtained from the two models and which model is preferable for predicting insolvency?*

2. Indicators

Indexes' selection was carried out by the CNDCEC[2] with the collaboration of the central financial statements of Cerved (http://press-magazine.it/wp-content/uploads/2019/10/codice-della-crisi_definizioni-indici_ott-2019.pdf). The objective was to identify a combination of indicators that were representative of income, equity or financial imbalance.

There were different critical thresholds, depending on different sectors to which the companies belonged. Specific alert indices have been introduced. The early warning indicators chosen by the CNDCEC and the thresholds for the manufacturing and industry business:

Financial expenses / Revenues from sales and services < 3,0%

Net equity / Total balance sheet assets > 7,6%

(Current assets + Cash and cash equivalents) / Short-term liability > 93,7%

Cash flow / Total balance sheet assets > 0,5%

Tax and social security debts / Total balance sheet assets < 4,9%

3. Altman's Z score model

Altman's Z score model is a best-known predictive model. It is based on a linear relationship that considers two society groups of equal size and five variables for each subject in the sample. Altman identifies five financial statement indices representative of the liquidity, profitability, debt and solvency of companies. Subsequently, the model was adapted to small and medium-sized enterprises and the Z' score model (Altman, 1993) was developed which is configured, in analytical terms, with the following linear function:

$$Z' = 0{,}717\ X_1 + 0{,}847\ X_2 + 3{,}107\ X_3 + 0{,}420\ X_4 + 0{,}998\ X_5$$

where the parameters represent: X_1: Working capital / Invested capital; X_2: Retained earnings / Invested capital; X_3: EBIT / Invested capital; X_4: Book capital of equity / Total debts; X_5: Sales / Invested capital. The Altman's Z score identifies three classification areas: Insolvency zone ($Z' < 1{,}23$); Uncertainty zone ($1{,}23 < Z' < 2{,}90$), and No risk zone ($Z' > 2{,}90$).

4. The dataset

The dataset includes data about 27,000 small and medium-sized enterprises (SMEs) headquartered in the Italian province of Brescia. It was downloaded by the AIDA dataset, following the recommendation of 6 May 2003 by the European Union Commission. The sample was determined with some mandatory requirements, such as the headquarters in the province of Brescia and available financial statements concerning 2022.

Additional sample requirements were used to select only the small and medium-sized enterprises as defined by the European Union Commission (Istat, 2022).

These additional requirements consider the enterprises with two of these three thresholds in 2022 financial statements: 1. Total production value (A) < 50.000.000 EUR; 2. Employees < 250;

[2] Consiglio Nazionale dei Dottori Commercialisti ed Esperti Contabili, *Crisi d'impresa – Gli indici dell'allerta, 20/10/2019*.

and Total balance sheet assets < 43.000.000 EUR. Each record considers economic and financial data from the 2022 financial statements published by the Chamber of Commerce of Brescia.

Firstly, the dataset calculates net equity > 0, a prejudicial indicator of the going concern of each SME. Some of the SMEs' indicators are not available because of the simplified Italian balance sheets required for the SMEs. Therefore, it became necessary to consider only the first four indicators for the data analysis and the last results. The Z' Score index is calculated considering the same previous dataset. Using this index, the companies are classified into three different areas: insolvency zone, uncertainty zone and no-risk zone.

We proceed by comparing the results determined by the indicators and the Z' score index of the 2022 financial statements with: Failed companies list from the *Court Law of Brescia*[3]; and Inactive or ceased companies list from the *Chamber of Commerce of Brescia*. Furthermore, we compared the results of the indicators with the results obtained with the Z' score index. The comparison aims to analyze the possible connection between the two results obtained and indicates which model is preferable to predict insolvency.

5. Results

Table 1 shows the results of the prejudicial indicator (net equity > 0) and the following five indicators. Data show that the prejudicial indicator is out of range for 2.345 companies. According to Italian Law, these companies (about 8,71% of the total sample) must begin the early warning procedure with the Chamber of Commerce. The other companies (24.582) must verify how many indicators are out of range. The results of the test are shown in Table 2.

Indicator	*Total Companies in the sample* (with data available) (A)	*Number of out-of-range Companies* (B)	*% of the sample* (B/A)
Net equity >0	26.927	2.345	8,71
Financial expenses/Revenues from sales and services < 3.0%	23.537	3.112	13,22
Net equity/Total balance sheet assets > 7.6%	26.793	5.109	19,07
(Current assets + Cash and cash equivalents)/ Short term liabilities > 93.70%	26.739	5.930	22,18
Cash flow/Total balance sheet assets > 0.5%	26.873	6.306	23,47
Tax and social security debts / Total balance sheet assets < 4.9%	12.603	3.869	30,70

Table 1 *Result of the analyzed indicators*

Number of out-of-range indicators	*Companies* (A)	*% of the sample* (A/24.582)
0	12.055	49,04
1	8.191	33,32
2	3.356	13,65
3	864	3,51
4	115	0,47
5	1	0,00

Table 2 *Out of range indicators*

The 116 SMEs with four or five out-of-range indicators must begin the early warning procedure with the chamber of commerce. With the application of the new Italian Law, 9,14% of the total

[3] The data includes the failures from 01/01/2020 to 31/10/2022 declared by the Court of Law of Brescia

sample must begin the early warning procedure. Table 3 shows the results of the Z' score index. With a median Z' score about of 1,95 (in the uncertainty zone) and an average Z' score of about 0,71 (in the insolvency zone). With the application of the Z' score index, 30,89% of the total sample has a serious insolvency risk.

Indicator	*Number of companies* (A)	*% of the sample* (A/26.927)
Insolvency zone (Z' <1,23)	8.319	30,89
Uncertainty zone (1,23 < Z' < 2,90)	10.987	40,80
No risk zone (Z' > 2,90)	7.621	28,31

Table 3 *Result of the Z' score index*

To compute how many companies have actually ceased or failed, we compared this data with the list of failed companies of the Court of Law of Brescia, the list of inactive or ceased companies from the Chamber of Commerce of Brescia, and the companies declared failed by the Court of Law of Brescia between 01/11/2022 and 31/05/2024 were about 231. 50 of these are included in our sample and the results are shown in Tables 4 and 5.

Number of out-of-range indicators	*Companies* (A)	*% of the sample* (A/50)
0	3	6
1	5	10
2	3	6
3	11	22
4	10	20
5	13	26
5 + net equity <0	5	10

Table 4 *Failed companies with the out-of-range indicators*

Z' Score	*Companies* (A)	*% of the sample* (A/50)
Insolvency Zone (Z' <1,23)	40	80
Uncertainty Zone (1,23 < Z' < 2,90)	7	14
No risk Zone (Z' > 2,90)	3	6

Table 5 *Failed companies with the Z' Score index*

With the sample of the SMEs in Brescia, we can analyze the inactive or ceased companies in 2022, 2023 and 01/01/2024-31/05/2024 and their out-of-range indicators and Z' score with Tables 6 and 7. The inactive or ceased companies included in our sample were 1.238.

Number of out-of-range indicators	*Companies* (A)	*% of the sample* (A/1.238)
0	140	11,31
1	326	26,33
2	142	11,47
3	165	13,33
4	356	28,76
5	101	8,16
5 + net equity <0	8	0,65

Table 6 *Inactive or ceased companies with the out-of-range indicators*

Z' score	*Companies* (A)	*% of the sample* (A/1,238)
Insolvency zone (Z' <1,23)	830	67,04
Uncertainty zone (1,23 < Z' < 2,90)	190	15,35
No risk zone (Z' > 2,90)	218	17,61

Table 7 *Inactive or ceased companies with the Z' Score index*

Data shows more out-of-range indicators and insolvency zone in the list regarding failed, inactive or ceased companies, compared to the total sample. Data are also shown in Tables 8 and 9.

Number of out of range indicators	*% of the total sample* (26,927)	*% of the failed, inactive or ceased Companies (1,288)*	*Difference*
0	44,77	11,10	-33,67
1	30,42	25,70	-4,72
2	12,46	11,26	-1,20
3	3,21	13,66	+10,45
4	0,43	28,42	+27,99
5 + net equity <0	8,71	9,86	+1,15

Table 8 *Comparison out-of-range indicators total sample – failed, inactive or ceased companies*

Z' Score	*% of the total sample* (26,927)	*% of the failed, inactive or ceased Companies (1,288)*	*Difference*
Insolvency Zone (Z' <1,23)	30,89	67,55	-36,66
Uncertainty Zone (1,23 < Z' < 2,90)	40,80	15,30	+25,50
No risk Zone (Z' > 2,90)	28,31	17,15	+11,16

Table 9 *Inactive or ceased companies with the Z' Score index*

6. Conclusions

This paper analysed the 2022 financial statements of the SMEs in Brescia with the list of failed companies of the Court of Law of Brescia and the inactive or ceased companies by the Chamber of Commerce of Brescia.

In the earlier study, through a dataset with about 20.000 SMEs with headquarters in Brescia Province, Italy, our previous research tried to verify if the indicators established by the CNDCEC are the right tools to predict the companies' failure. None of the companies in the sample failed in 2020, 2021 and 2022.

Are the indicators the right tool to predict the company's failure?

The new data shows that this model helps to predict the crisis state, but it is not enough: in the 50 failed companies included in the sample, 22% of them have less than 3 indicators out-of-range.

Instead, the Z' score index is more selective, and, in our research, it results to be more appropriate: The no-risk zone for the failed companies is only 6% of the total failed companies. If we compare, however, the inactive or ceased companies, about 49,11% of them have less than 3 indicators out-of-range. The no-risk zone is about 17,61%. Therefore, we can see that the Z' score model is more selective and requires better financial stability to enter the no-risk zone instead of the indicators model.

Maybe one of the possible causes could be that the indicators models are referred to as an obligation for the administrators with important responsibilities in case of non-compliance. The Z' score model, instead, is referred to as economics and financials analysis.

In our opinion, introducing the indicators' model, the companies have to improve their

budgeting and management control skill and this is an opportunity for the Italians SMEs to reach a better decision-making process. By analysing the data, these models can be an important opportunity to develop predictions analysis in the Italian SMEs and it may be useful for business strategy and management.

The research could be extended by analysing other Italian provinces and with a multi-year dataset or by introducing a new and better indicator model created by the sample, possibly linked with the Z' score model maybe using machine learning techniques.

References

Alberici A. (1975). *Analisi dei bilanci e previsione delle insolvenze: affidabilità bancaria e informativa del mercato mobiliare*, Isedi, Milano.

Altman E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, 23(4), pp. 589-609.

Altman E.I (1993). *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting & Avoiding Distress and Profiting from Bankruptcy*, Wiley, New York.

Altman E.I. (2002). Revisiting credit scoring models in a Basel 2 environment (May 2002). *NYU Working Paper No. S-FI-02-11*, SSRN: https://ssrn.com/abstract=1298829.

Altman, E.I., Falini A. Danovi A (2013). Z-score models' application to Italian companies subject to extraordinary administration. *Bancaria* No. 04-2013, SSRN: https://ssrn.com/abstract=2275390

Altman E.I., Hotchkiss E (1995), *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting and Avoiding Distress and Profiting from Bankruptcy,* Wiley, New York.

Bottani P., Cipriani L., Serao F. (2004). Il modello di analisi Z-score applicate alle PMI, *Amministrazione & Finanza,* 1, pp. 50-53.

Brentari, E., Ermici, G., Zanetti, A. (2024), The new preventive alert mechanism in the Italian crisis code: A statistical approach for predictive insolvency analysis. In *Statistics, Technology and Data Science for Economic and Social Development, Book of Short Papers of the ASA Bologna Conference – Supplement to Volume 35/3* – Available at doi.org/10.26398/asaproc.0012.

Chye K.H., Kee L.C. (2004). Going concern prediction using data mining techniques, *Managerial Auditing Journal*, 19(3), pp. 462-476.

Consiglio Nazionale dei Dottori Commercialisti ed Esperti Contabili, *Crisi d'impresa – Gli indici dell'allerta, 20/10/2019*. http://press-magazine.it/wp-content/uploads/2019/10/codice-della-crisi_definizioni-indici_ott-2019.pdf.

International Standard on Auditing 570 (Revised) (2016). *Going Concern*. https://www.ifac.org/_flysystem/azure-private/publications/files/ISA-570-(Revised).pdf.

ISTAT (2022), *Annuario statistico italiano 2022*. https://www.istat.it/storage/ASI/2022/capitoli/C14.pdf. ISBN 978-88-458-2090-86, Chapter 14.

ISTAT (2023). *Registrazione e fallimenti di imprese, I trimestre 2023.* https://www.istat.it/it/files//2023/05/STATFLASH_FALLIMENTI-IMPRESE_I-TRIM-2023.pdf.

Italian Accounting Organism (2018). *OIC 11, Finalità e postulati del bilancio d'esercizio*, marzo 2018. https://www.fondazioneoic.eu/wp-content/uploads/2011/02/2018-03-OIC-11.pdf.

Sterling R.R. (1968). The going concern: An examination, *The Accounting Review*, 43(3), pp. 481-502.

doi.org/10.26398/asaproc.0086

An innovative point of view for statistical dissemination to prevent road accidents using the 'collision matrix' [1]

Marco Broccoli[a], Silvia Bruzzone[b]

[a] Istat - Italian National Institute of Statistics - Directorate for Methodology and Statistical Process Design
[b]Istat - Italian National Institute of Statistics - Directorate for Social Statistics and Welfare

1. Introduction

Road accidents pose a complex issue that requires comprehensive data analysis through the use of new indicators and innovative approaches, moving beyond traditional methods. To study the various features of road users and their interactions, stakeholders must relate specific risk elements to the connections between users and vehicles. Currently, an information bias persists due to the choice of inappropriate denominators for calculating road accident rates. Analysts often use the resident population or vehicle fleet as common proxies for individuals exposed to risk, but these proxies do not always provide appropriate solutions. The new collision matrices address this gap in road accident analysis and offer a fresh perspective for investigation.

This study also defines key elements for measuring road safety performance with the support of Key Performance Indicators (KPIs) outlined in the 2030 Agenda for European Union countries. The new method applies a multidimensional approach to road safety, focusing on roads, vehicles, and human behaviours as crucial components.

In addition to traditional mortality and injury rates, and in relation to road length, the new collision matrices introduce an innovative aspect to the project, as showcased in the Istat dissemination product. This precise technique connects fatalities (within 30 days) or injuries from road accidents with drivers, passengers, pedestrians, and the types of vehicles involved in the collisions

2. The road accidents resulting in death or injuries survey

The survey on road accidents resulting in death (within 30 day) or injury carried out by the Italian National Institute of Statistics (Istat), with the cooperation of ACI (Automobile Club of Italy) and other local organisations, is an exhaustive and monthly based data collection, included in the National Statistical Programme. The survey collects all road accidents involving at least a vehicle circulating on the national road net, resulting in death or injury and documented by a Police authority (Council Decision n. 704 of 1993, of 30 November 1993)[2].

The detection unit is the single road accident resulting in death or injury; all information is referred to the period when the accident occurred.

As regards the data flow, Istat adopted a flexible model, through the subscription of a memorandum of understanding or special agreements signed with regions (NUTS2 level) and provinces (NUTS3 level). Main information collected are date, time and location of the accident, type of road and surface, signals, weather conditions, type of accident (collision, investment, etc.), type of vehicles involved, causes of the accident.

[1] Marco Broccoli and Silvia Bruzzone edited Section 1; Silvia Bruzzone edited sections 2 and 3; Marco Broccoli edited Sections 4 and 5, Marco Broccoli and Silvia Bruzzone edited Section 6.
[2] Link to: Decision EC 704/93

3. The road accidents in Italy and in Europe: An overview on recent data

During 2023, a stabilization in mobility occurs, if compared to 2022, a year in which there was a clear increase in noticed movements compared to the most acute phases of the pandemic. Concerning road accidents, 2023 shows a small improvement in the number of fatalities compared to the previous year; however, there is an upward trend in accidents and injuries, although slightly.

In 2023, we count 3,039 deaths in road accidents in Italy (-3.8% compared to the previous year), 224,634 injuries (+0.5%), and 166,525 road accidents (+0.4%). The values are slightly higher than 2022 for accidents and injuries, but lower for fatalities. There is still a decrease compared to 2019 for accidents, fatalities, and injuries (respectively -3.3%, -4.2%, and -6.9% (Figure 1 and Table 1).

In 2023, fatalities increased for drivers of e-scooters, bicycles, and electric bicycles, they are stable for pedestrians, and decreased for other users. There were 1,332 fatalities among car occupants (-3.1%), 734 among motorcyclists (-6.0%), 68 among moped riders (-2.9%), 485 among pedestrians (0.0%). Among truck occupants, 112 deaths were recorded (-32.5%), while for bicycles and electric bicycles, the fatalities were 212, an increase compared to 2022 when there were 205 (+3.4%). Injuries also increased among users of electric scooters (counted since 2020): road accidents involving them rose from 2,929 in 2022 to 3,365 in 2023, injuries from 2,787 to 3,195, while deaths (within 30 days) were 21 (in 2022 there were 16).

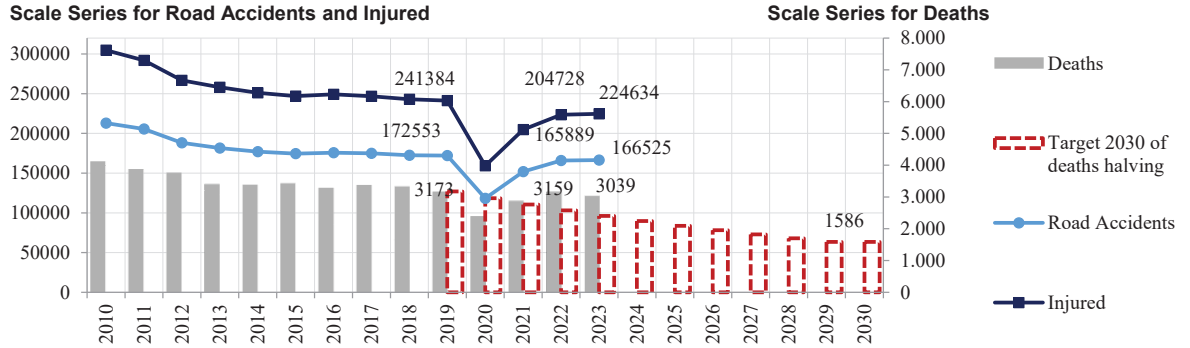
On urban roads, motorways and rural roads, road accidents and injuries are slightly up or stable compared to 2022, with values still all falling compared to the reference year 2019. The number of fatalities fell in comparison with 2022 in all road areas, with a substantial drop on motorways (-19.0 per cent); on urban roads, the decrease was 0.3 per cent, on rural roads 3.9 per cent.

Distraction, failure to yield right of way and speeding are confirmed as the most frequent types of incorrect driving behaviour. The three groups together account for 36.5% of cases (80,057), a value that has remained stable over time.

Driving too fast is the most sanctioned behaviour after prohibited stop, parking, and accounts for 37% of all traffic offences. Sanctions for failure to use seat belts and child restraint systems fell slightly, while those for failure to wear a helmet rose sharply. The number of penalties for improper use of devices in the car remains high, and penalties for driving under the influence of alcohol are on the rise, especially the high proportion from local police forces.

The car market shows a growth in 2023: first registrations of passenger's cars increased by 18.4 % compared to 2022. On the motorway network, the annual average vehicle journeys show a growth of 3.8% compared to 2022 and an increase compared to 2019, recording a maximum, in absolute terms for total journeys, which exceed 86.7 billion vehicles per km.

Figure 1 – *Road Accidents resulting in death or injury, deaths and injured. Years 2010-2023. (absolute values)*



Source: Istat Survey on road accidents resulting in death or injury.

4. Collision matrices: The methodology used

The "collision matrices" represent the original element of the project, calculated by road type and other crucial features. The matrix structure consists in an accurate technique, aimed to put in connections all persons killed (within 30 days) or injured in road accidents, drivers and passengers, or pedestrian and the vehicle type, responsible of the collision during the crash.

The road accident indicators based on collision matrices make it possible to outline risk profiles for road users and to highlight differences by type of road, urban and rural, by geographical zone of accident event and by gender and age of users involved.

The project was born following a format used by the European Commission[3], implementing a national variation. The Italian research provides, in fact, an updated and enriched version, in respect to the European Commission proposal.

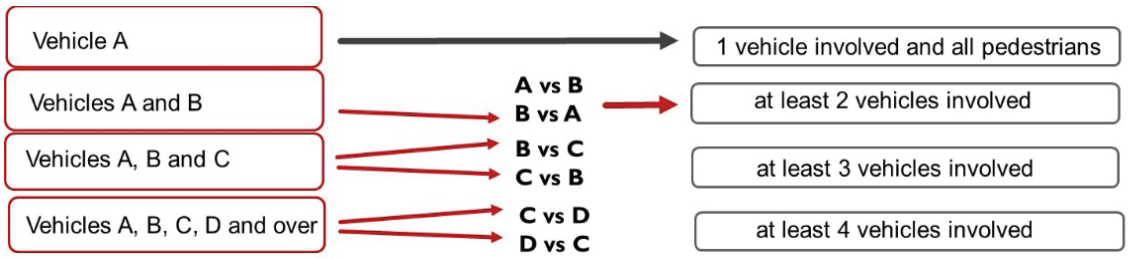
The collision matrix proposed by EU include data with coverage of fatalities in single-vehicle crashes and crashes involving one or more traffic units. For most fatal crashes, only one other vehicle is involved in the crash. For multi-vehicle crashes, the "main vehicle" is the heaviest of the vehicles involved, as this tends to be responsible for the most serious consequences. As results, the figures in each column likely underestimate the number of cases a particular vehicle was involved in a crash (Source: EU CARE Database). We add that, considering this approximation, a bias could be introduced, linked to the combinations between vehicles that do not collide each other and involved in the same accident. Nevertheless, we believe the computing is a good "proxy" for an exhaustive analysis. Accidents with more than 2 vehicles involved not exceed 10%.

Concerning the Italian version, the values within the cells of the matrix represent, as an added value, a partition of total cases (killed or injured), developing an algorithm combining and permuting all vehicles involved in the accidents and all the crashing interaction between vehicles and other vehicles or pedestrians. Another customisation implemented by the Italian researchers, in the matrix Italian version is the introduction of two new types of vehicles: E-bike and E-scooters.

The algorithm for building the collision matrix follows an iterative process with distinct steps. First, it analyzes road accidents involving a single vehicle, including incidents with pedestrians. Next, it examines accidents with at least two vehicles by comparing the first two vehicles (A vs B and B vs A). Finally, for accidents involving three or more vehicles, the process compares additional vehicles (B vs C and C vs B), ensuring that comparisons involving the first vehicle (A) have already been addressed. This sequential approach ensures a comprehensive analysis of all vehicle interactions.

The procedure of extraction from a star-schema data warehouse, uses nested SQL queries to build a result set, combining facts and dimensions with cartesian product of metadata for the dynamically generating pivot tables in Excel. The next steps follow the extracting process (Figure 2).

Figure 2 - *General scheme of the iterative procedure for collision matrices processing*



Source: Istat processing

[3] European Commission: 2023 figures show stalling progress in reducing road fatalities in too many countries - European Commission (europa.eu)

5. Collision matrices: Main results

The collision matrices produced contain data for a set of years, in time series, considering the main benchmarks years for the last road safety decades (2001, 2010 and 2019) and the most recent data too (2022 and 2023)[4].

They refer to fatalities and injuries for drivers, passengers or pedestrians and detailed for urban and rural areas. The collision matrix includes, as an innovative analysis element, the information on the mortality or harmfulness risk (as a proxy) for the different user's categories. The tables contain the absolute number for the cross-referencing of the matrix cells, the row percentage values and the column percentage values. The reading key, by the percentage values information too, is useful to understand immediately the weight of each category out of the total.

The analysis of the main and most recent data for 2023 shows interesting results. Concerning mortality data, for the most vulnerable users, the drivers of electric bicycles and scooters collide mainly against passenger cars, or they are involved in single vehicles accidents. Powered two-wheelers have a high number of fatal accidents in collisions with passenger cars, light commercial vehicles and as single vehicles too. Pedestrians present a higher mortality risk than other users when they collide with passenger cars or with heavy good vehicles. Passenger's cars user crash mainly with other cars or as single vehicles. Heavy good vehicles collide against other heavy good vehicles, passenger's cars, and in single vehicle accidents (Tables 1, 2 and 3).

Table 1 - *Collision matrix for road accidents fatalities by road user and other vehicles involved. Year 2023 (absolute values)*

	IN COLLISION WITH											
FATALITIES BY USER TYPE	**Bicycles**	**E-bikes**	**E-scooters**	**Mopeds**	**Motorbikes**	**Passengers' cars**	**Lorries (< 3,5 t)**	**Heavy Good Vehicles**	**Buses or Coach**	**Other vehicles**	**No other vehicle involved**	**Total**
Pedestrians	4	2	0	4	33	353	37	28	14	10	0	**485**
Bicycles	6	0	0	0	9	116	18	14	4	7	26	**200**
E-bikes	0	0	0	1	1	7	1	0	0	0	2	**12**
E-scooters	0	0	0	0	2	6	0	1	0	0	12	**21**
Mopeds	0	0	0	3	2	29	4	1	0	4	25	**68**
Motorbikes	2	0	1	0	33	375	50	23	6	4	240	**734**
Passengers' cars	1	0	0	4	4	532	68	123	15	21	564	**1,332**
Lorries (< 3,5 t)	0	0	0	0	0	9	5	8	2	0	18	**42**
Heavy good vehicles (> 3,5 t)	0	0	0	0	0	11	2	27	5	0	25	**70**
Buses or coach	0	0	0	0	0	0	0	5	0	0	23	**28**
Other vehicles	0	0	0	0	1	20	3	9	0	0	14	**47**
Total	**13**	**2**	**1**	**12**	**85**	**1,458**	**188**	**239**	**46**	**46**	**949**	**3,039**

[4] Istat Road accidents in Italy. Year 2023 (Appendices) https://www.istat.it/en/press-release/road-accidents-2023/

Table 2 - *Collision matrix for road accidents fatalities by road user and other vehicles involved. Year 2023 (Row percentage values)*

	IN COLLISION WITH											
FATALITIES BY USER TYPE	**Bicycles**	**E-bikes**	**E-scooters**	**Mopeds**	**Motorbikes**	**Passengers' cars**	**Lorries (< 3,5 t)**	**Heavy Good Vehicles**	**Buses or Coach**	**Other vehicles**	**No other vehicle involved**	**Total**
Pedestrians	0.8	0.4	0.0	0.8	6.8	73.0	7.6	5.6	2.9	2.1	0.0	100.0
Bicycles	3.0	0.0	0.0	0.0	4.5	58.0	8.5	7.5	2.0	3.5	13.0	100.0
E-bikes	0.0	0.0	0.0	8.3	8.3	58.3	8.3	0.0	0.0	0.0	16.8	100.0
E-scooters	0.0	0.0	0.0	0.0	9.5	28.6	0.0	4.8	0.0	0.0	57.1	100.0
Mopeds	0.0	0.0	0.0	4.4	2.9	42.6	5.9	1.5	0.0	5.9	36.8	100.0
Motorbikes	0.3	0.0	0.1	0.0	4.5	51.1	6.8	3.1	0.8	0.5	32.8	100.0
Passenger's cars	0.1	0.0	0.0	0.3	0.3	39.9	5.1	9.2	1.1	1.6	42.4	100.0
Lorries (< 3,5 t)	0.0	0.0	0.0	0.0	0.0	21.4	11.9	19.0	4.8	0.0	42.9	100.0
Heavy Good Vehicles (> 3,5 t)	0.0	0.0	0.0	0.0	0.0	15.7	2.9	38.6	7.1	0.0	35.7	100.0
Buses or Coach	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.9	0.0	0.0	82.1	100.0
Other vehicles	0.0	0.0	0.0	0.0	2.1	42.6	6.4	19.1	0.0	0.0	29.8	100.0
Total	0.4	0.1	0.0	0.4	2.8	48.0	6.2	7.9	1.5	1.5	31.2	100.0

Table 3 - *Collision matrix for road accidents fatalities by road user and other vehicles involved. Year 2023 (Column percentage values)*

	IN COLLISION WITH											
FATALITIES BY USER TYPE	**Bicycles**	**E-bikes**	**E-scooters**	**Mopeds**	**Motorbikes**	**Passengers' cars**	**Lorries (< 3,5 t)**	**Heavy Good Vehicles**	**Buses or Coach**	**Other vehicles**	**No other vehicle involved**	**Total**
Pedestrians	30.8	100.0	0.0	33.3	38.8	24.2	19.7	11.7	30.4	21.7	0.0	16.0
Bicycles	46.2	0.0	0.0	0.0	10.6	8.0	9.6	5.9	8.7	15.2	2.7	6.6
E-bikes	0.0	0.0	0.0	8.3	1.2	0.5	0.5	0.0	0.0	0.0	0.2	0.4
E-scooters	0.0	0.0	0.0	0.0	2.4	0.4	0.0	0.4	0.0	0.0	1.3	0.7
Mopeds	0.0	0.0	0.0	25.0	2.4	2.0	2.1	0.4	0.0	8.7	2.6	2.2
Motorbikes	15.4	0.0	100.0	0.0	38.7	25.7	26.6	9.6	13.0	8.7	25.3	24.2
Passengers' cars	7.6	0.0	0.0	33.2	4.7	36.4	36.1	51.5	32.7	45.7	59.5	43.8
Lorries (< 3,5 t)	0.0	0.0	0.0	0.0	0.0	0.6	2.7	3.3	4.3	0.0	1.9	1.4
Heavy Good Vehicles (> 3,5 t)	0.0	0.0	0.0	0.0	0.0	0.8	1.1	11.3	10.9	0.0	2.6	2.3
Buses or Coach	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	2.4	0.9
Other vehicles	0.0	0.0	0.0	0.0	1.2	1.4	1.6	3.8	0.0	0.0	1.5	1.5
Total	100	100	100	100	100	100	100	100	100	100	100	100

As regards the vehicles involved in the collision, passengers' cars drivers are responsible, in total, mainly for other car occupants' deaths (36.4%), motorcyclists' fatalities (25.7%), and pedestrians (24.2%, 39% in urban area). Motorcyclists, instead, kill mainly, pedestrians (38.8%) and other motorcyclists (38.7%) (Table 4).

The analysis of collision matrices data for fatalities within urban areas shows a slightly different profile for users and the type of vehicles in collision with them enter during the crash, if compared with the total data or with rural roads information.

Cyclists, in fact, reach a percentage of collision with passengers' cars of 64.2%, while the value is 58% as whole and 51.1% on rural roads. Concerning e-scooters, the collision is more frequent within urban roads (14 deaths in 2023) than outside the built-up areas (7 deaths in 2023). The collision outside the cities is mainly with passengers' cars (57.1%); inside the urban roads, the percentage is

highest for accidents as single vehicles due to the loosing of control of the means of transportation (78.6%). A possible cause could be the condition of the surface of the road or the distraction or the misuse of the vehicle, without the respect of the rules of use. For motorcyclists, ever for crashes with passengers' cars, the percentage of collision rises to 54.1% from 51.1% as whole and 48.3 outside built up areas. Injuries matrices show similar results and profiles within and outside urban areas.
Finally, an interesting comparison consists in the time series collision matrices analysis too. One of the most evidence emerged, for instance, is the increasing of the percentage of single vehicles accidents, within built up areas. The percentage rises, in fact, since 38.3% in 2001 (Table 9) to 51.9% in 2023 (Table 6). In addition, drivers of lorries, classified as light commercial vehicles, shows a strong increase of single vehicle accidents, from 34.8% to 55.6%.

The specific increase could partly result from distractions caused by using smartphones or devices while driving, along with bad road conditions or speeding. Some studies (Matthews et al., 2019) deal with the strong connection between fatigue and distraction driving too, or between other activities done while driving, such as eating or drinking, and texting (Choudhary et al., 2017; Gariazzo et al., 2018)). The rise in distractions often leads to road accidents for lorry drivers as well, especially in recent years, with the growing spread of e-commerce and delivery networks.

6. Conclusions

Collision matrices represent a new tool for analysing road accident data, allowing for the study of connections between road users—whether injured, killed, or hurt—and the vehicles they collided with, which are often responsible for the accident. The scientific community particularly values the analysis using two-way tables because it explains data clearly and straightforwardly, filling an existing information gap in the available tools. The generalization of collision matrices for the Italian model introduces two innovations: the inclusion of electric scooters and electric bicycles, and the calculation that allows for a data partitioning in the table, with marginal distributions that include all recorded cases of deaths and injuries, not just those involving the first two vehicles, as proposed by the European Commission.

References

Choudhary, P., Velaga, N.R. (2017) Modelling driver distraction effects due to mobile phone use on reaction time, *Transportation Research Part C: Emerging Technologies,* 77, pp. 351-365.

ETSC (European Tranport Safety Council) (2024). Ranking EU progress on road safety 18th Road Safety Performance Index (PIN) Report (2024).

European Commission (2024) (Collision Matrices) 2023 figures show stalling progress in reducing road fatalities in too many countries - European Commission (europa.eu)

Gariazzo, C., Stafoggia, M., Bruzzone, S., Pelliccioni, A., Forastiere, F. (2018). Association between mobile phone traffic volume and road crash fatalities: A population-based case-crossover study. *Accident; Analysis and Prevention, 115,* pp. 25-33.

Istat (2024), *Road Accidents in Italy. Year 2023 (2024),* (Appendices), Road accidents 2023 – Istat.

Matthews G., Neubauer C., Saxby D.J., Wohleber R.W., Lin J. (2019), Dangerous intersections? A review of studies of fatigue and distraction in the automated vehicle, *Accident Analysis and Prevention*, 126, pp. 85-94.

doi.org/10.26398/asaproc.0087

A new algorithm for the correction and control of road accidents data based on string similarities

Silvia Bruzzone [b], Luca Valentino [a], Diego Moretti [a], Massimo Tavanti [b], Francesca De Blasi [c]

[a] Istat - Italian National Institute of Statistics - Directorate for Methodology and Statistical Process Design

[b] Istat - Italian National Institute of Statistics - Directorate for Social Statistics and Welfare

[c] ACI Informatica - Directorate Demand & Service ACI

1. Introduction

The Italian National Institute of Statistics (Istat) collects road accidents data in Italy, focusing on accidents causing injured or deaths within 30 days after the event. The data collection involves cooperation with local police, the Automobile Club d'Italia (ACI), regions, and autonomous provinces, provinces and municipalities, following specific agreements. The survey covers all road accidents reported by police resulting in death or injury, excluding crashes with only vehicles damages, those in non-public areas, and accidents without vehicles involvement. The results have been published on a regular basis since 1952.

Since 1991, Istat data collection complies with international definitions, ensuring consistent and accurate statistics. Information gathered includes the date, time and location of accidents; type of road; weather conditions, involved vehicles and information on drivers, passengers, and pedestrians. Istat devotes great deal of attention to quality of source by strict checks on data and subsequent correction procedures. One of the fundamental step in the quality process is also the de-duplication of road accidents occurrences. The presence of duplicate records has, obviously, a deep impact on the results of the survey and, consequently, on the evaluation of current national road safety policies.

2. Edit, imputation and deterministic rule for the identification of duplicates

The unit of measurement is the road accidents resulting in death or injuries. The survey refers to the date when the accident occurred. As usually happens for this type of sources, also road accidents data contains errors and missing values, than it needs to be processed beforehand with correction procedures. In particular, in this case the source may contain duplicates of records too. In fact, both Police and Carabinieri can intervene in the same scene of accident and verbalise the same accident. Sometimes for different reasons the same Police station can record doubling the same event too. These frequently remain as duplicated in their own databases.

The software used to process the probabilistic check and corrections is Con.Cor.D. (Controllo e Correzione Dati) an open source tool developed by Istat, current version 2.2 is developed in Java (2014)[1]. The software, together with a complex procedure for deterministic imputation, assure the accuracy for the probabilistic error correction methods. The software has been used since 2000 and can manage various data correction scenarios, based on predefined rules and statistical techniques. An important role for the data correction procedure assume the deduplication record processp

Until data referred to 2021, to identify and eliminate the double records in the final data file, a complex, SAS (a package aimed to a Statistical Analysis System) based and deterministic procedure has

[1]Software and user manual for Concord java version 2.2 is available on Istat website at url: https://www.istat.it/en/classifications-and-tools/methods-and-software-of-the-statistical-process/process-phase/detection-and-treatment-of-measurement-errors-and-imputation-of-partial-non-responses/concordjava/

been implemented at Istat.

The deterministic procedure is based on the use of "key variables", obtained by concatenating the codes of multiple variables, crucial to subsequent check of the duplicate records, i.e., records with the same composition of codes in the above-mentioned variables.

After the key variable construction, the procedure based on "FIRST and LAST" method, by SAS, is applied. The procedure assigns, as a first step, a binary code: 0 or 1, when the selected variables are filled in or not and if a similarity is detected between records examined by pairs. The following sequences of 0 and 1 indicate whether the record being analysed is original or has duplicates[2]. The key variables used in this step are related to the date of accident (month, day and hour) and to the vehicles involved (type of vehicles and number plate).

The procedure is iterative and replicated more times, varying the composition of the key variable and applying a fine-tuning method to arrive to an ever better result. After the SAS procedure processing, every year, in average, about 100 duplicated records are excluded from the final dataset.

During the data control and correction data phase, since 2022 year of reference data, to additionally improve the deduplication system approach, Istat decided to implement a probabilistic module too.

After the initial filtering of duplicates, by means SAS and the deterministic approach, in fact, a deduplication procedure based on RELAIS (REcord Linkage At IStat), package implemented at Istat, was also applied.

3. The record linkage problem and RELAIS

After the deterministic procedure (§ 2), to detect the residual duplicates record we adopt a probabilistic approach with a task solving as a record linkage case. Record linkage aims to find records that refer at the same entity in the real world. If the records are part of separate data sets it completes a process of data integration, if the records are part of a single data set, as in our case, it resolves a problem of de-duplication.

Istat developed RELAIS (REcord Linkage At IStat), a software for record linkage and data deduplication (Cibella et al. 2015) [3]. It is an open-source toolkit, addressed to different phases of record linkage, including pre-processing data, search space selection, comparison and decision modelling. It supports multiple methodologies, as well as deterministic and probabilistic models, like the Fellegi-Sunter algorithm.

The tool offers several string comparison metrics, for instance Levenshtein, Jaro, Q-Grams, and Jaro-Winkler ensuring robust data matching. Moreover, RELAIS make available advanced techniques such as SimHash that are not available in the other most frequently used tools.

The variables we intend to use in this probabilistic step for duplicates searching are the name of people involved in the road accident: deaths (within 30 days) and injured. The data schema counts 24 columns for these, name and surname for six deaths and six injured.

It is crucial to note that these fields present some problems that make them difficult to use. In particular, how shown in table 1 as example, we observe:

- Typographical errors
- Incorrect assign of name and surname (in particular for foreigners)
- Different order of people in the list
- People listed as "dead" in a record and listed as "injured" in the other

[2] Specifically:
If FIRST is 1 and LAST is 1, we are dealing with an original record for which there are no duplicate records.
If FIRST is 1 and LAST is 0, we are dealing with an original record for which there are subsequent duplicate records.
If FIRST is 0 and LAST is 0, we are dealing with a duplicate record.
If FIRST is 0 and LAST is 1, we are dealing with a duplicate record.

[3] Software and user manual for Relais version 3.1 is available on Istat website at url: https://www.istat.it/en/classifications-and-tools/methods-and-software-of-the-statistical-process/process-phase/data-integration/relais/

Table 1 – *Example of duplicated record with errors*

	Record 1			Record 2	
Death 1	Valentino	Luca		Luca	Valentino
Death 2	Moretti	Diego			
Injured 1	Bruzzone	Silvia		Ruzzone	Silvia
Injured 2	Tavanti	Massimo		De Blasi	Francesca
Injured 3	De Blasi	Francesca		Tavanti	Massimo
Injured 4				Moretti	Diego

Due to these problems is not profitable to make an exact comparison between the records in question. Therefore, the use of approximate metrics is preferred. Conversely, the utilisation of approximate similarity measures is more time-consuming, particularly when dealing with large datasets, as is the case here. So reducing the complexity of the problem is a necessary step that Relais provides in the phase named 'Search space reduction'. Only pairs of record selected in the search space are then compared.

In the proposed process of de-duplication we perform a search space reduction using the probabilistic methods SimHash that appears confident to trait the expected errors in field value. In particular, we apply the method to the concatenation of all involved people names (deaths and injured). This step is the really crucial one of the whole elaboration; the details of SimHash algorithm are better described in the next paragraph.

Once the problem is confined to a very limited number of record pairs, each pair is processed by comparing the names of the deaths (concatenated into a single string) and the injured (in a single string too) and the two together with the string similarity functions *HSim* (based on SimHash) also described in the next paragraph. If one of these three strings is found to be similar in records beyond a predefined threshold, the record pair is selected as duplicate.

Initially, the threshold is set to a highly restrictive value of 0.9. However, depending on the result of the specific process, it can be decreased just to a value of 0.8. The duplicates thus obtained are also subject to manual revision prior to the definitive deletion from source.

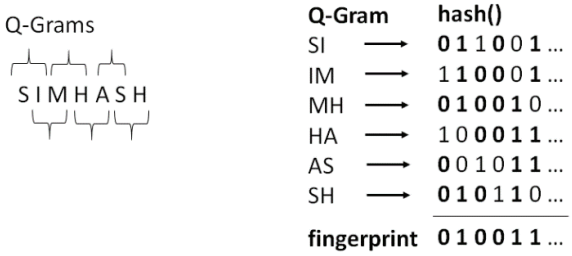
4. Methods

SimHash is a locality-sensitive hashing (LSH) technique used to finding near-neighbours in high dimensional space. Initially developed for the detection of duplicate texts (Charikar 2002), it has also proven effective for comparing short strings (Mancini et al. 2012). In the context of text comparison, the method is based on words, in the case of short strings (as implemented in RELAIS), it is based on q-grams, i.e. sub-strings composed of q consecutive characters. The method enables the comparison of a large number of strings by sorting appropriately the records and comparing only the closest instances in the sorted list. To ensure that the 'closest' status does not depend exclusively on a few particular characters (e.g. in alphabetical sorting it depends only on the first characters) the sorting step is repeated several times with different criteria. To achieve this objective in an efficient and straightforward manner, as will be shown, SimHash does not work directly with the strings but with the hash representation of these.

In the practice, the algorithm works as follows:

In first step SimHash calculate the hash representation of the strings (in the following fingerprints). The fingerprint is a binary vector of fixed dimension (128 in this case). The calculation of this, how said, is based on q-grams of the string. How shown in example of Figure 1, we consider the hash value of each q-gram and the synthesis of these compose the fingerprint of the string. The fingerprint is the vector where each values is the mode of corresponding values in the q-grams hash vectors.

Figure 1 - *Calculation of fingerprint for string 'Simhash'*



In formula, being $q_1,\ldots, q_n$ the *n* q-grams contained in the string *s*, we define the vector $\vec{h}_i$ for *i* from 1 to *n* as result of the hashing function

$$\vec{h}_i = \left(h_{i,1}, \ldots, h_{i,128}\right) = hash(q_i)$$

and $h_{i,j}$ may assume value of 1 or 0. Then the fingerprint *f* of string *s* is the vector

$$\vec{f} = f(s) = (f_1, \ldots, f_{128})$$

Where

$$f_j = mod_j(s) = \begin{cases} 1\ if\ \sum_{i=1}^{n} h_{i,j} > n/2 \\ 0\ if\ \sum_{i=1}^{n} h_{i,j} \leq n/2 \end{cases}$$

To compare *f* and *f'* fingerprints of strings *s* and *s'* we use the *Hamming distance* (*HD*).

The HD take as value the number of positions at which the corresponding element of *f* and *f'* are different.

As a consequence of the characteristics of the *hash()* function, it results that, fixed the index *j*, the value for j-component $h_{i,j}$ of vector $\vec{h}_i = hash(q_i)$ can be considered as a random number (between 0 and 1), this is independent from others $h_{x,j}$ and depends only on the q-gram q_i .

More in detail if q_i is a q-gram of *s*, q_i' a q-gram of *s'* and $\overrightarrow{h_i'} = hash(q_i')$, the probability that $h_{i,j} = h_{i,j}'$ is:

$$Pr\left(h_{i,j} = h_{i,j}'\right) = \begin{cases} 1\ \ if\ q_i' = q_i \\ 0.5\ \ if\ q_i' \neq q_i \end{cases}$$

Furthermore, given the definition of fingerprints *f* and *f'*, it is clear that the probability that $f_j = f_j'$ increases as increases the number of common q-grams is *s* and *s'*. Charikar (2002) shows that the *HD* between two fingerprints approximates the cosine distance between the q-grams of the corresponding strings. This is the typical metric based on q-grams used for string comparison (found in RELAIS with q=3 and name *3-Grams*). We can use the *HD* to decide about similarity of strings.

The sorting step is also performed using the fingerprints. This allows repeating different sorting using permutations of fingerprint. In effect, the entire fingerprint have the same dimension (128) and the *HD* is

invariant with permutations.

The SimHash method decreases the complexity of this phase from $O(n^2)$ to, approximately, $O(n)$ per sorting step. It can be reasonably assumed that all similar record pairs will end up near in some sorted list if an adequate number of rotations with different permutation are used.

The configuration parameters of the SimHash method are three: q (*q-grams* dimension), *threshold* (threshold on *HD* to discriminate if two fingerprints are "similar" or not) and *rotations* (the number of fingerprint permutations, and resulting sorting steps). When using short strings, the parameter q is generally selected between low values (2 or 3). This can be raised to a value of 4 or 5 if the average size of the strings is on higher (several tens of characters). In our case the strings are typically short than we use the default value recommended in Relais, i.e. q=2. Also for *threshold* we use the default value recommended in Relais, i.e. 40. In practice this is a threshold sufficiently restrictive to ensure a good reduction of candidate pairs. In fact we observe that using a value greater than 40 for *threshold* assess a significantly growth of candidate pairs while the final number of duplicates remains more or less unchanged. Finally, with regard to the rotations parameter, we observe that increasing the number of permutations increases the time used by the process but not always the number of similar pairs extracted; in our case, we found that four rotations are be sufficient for a reliable result.

The metric *HSim,* used in the final rule for duplicates selection, is also based on SimHash fingerprints and on *HD*. More in detail is the complementary to 1 of the *HD* normalized.

$$HSim(s_1, s_2) = 1 - \frac{HD(f_1, f_2)}{128}$$

Where f_1 and f_2 are the fingerprints of strings s_1 and s_2, and 128 is the maximum value for *HD.*

5. Scalability: The blocked SimHash

In case of very large datasets, even an optimized method as SimHash may suffer of a problem of insufficient resources. As example RELAIS performs the sorting steps in RAM and thus have a physical constraint on the space available for the fingerprints vector storing. Furthermore, the scalability of the method is difficult because the entire fingerprints set must be processed together. Other LSH methods, like MinHash, also suffer from this same problem.

A good scalable approximation of SimHash is the 'Blocked SimHash' method, which divides the data set into disjoint blocks using a blocking variable. In our case, since the police authorities are organised at provincial and municipality level, the province is a good variable to use for this purpose. For each block, the method is only applied to records from a province. RELAIS, which is not a scalable tool, processes the blocks sequentially; in a scalable system, the blocks can be processed in parallel.

6. Results and conclusions

During the data control and correction data phase, after the initial filtering of duplicates, by means SAS and a deterministic approach, the deduplication procedure based on RELAIS was also applied in two iterative distinct steps.

The process was fine-tuned using final 2023 data. The dataset consists of 163,317 records[4] relating to road accidents that occurred on all Italian territory. The search space created performing the *SimHash* method counts about 400,000 candidate pairs. As previously detailed, the final parameters of *SimHash* process are the following: q=2 (*q-grams* dimension), 4 sorting rotations and 40/128 elements as threshold for the *Hamming distance*.

[4] Excluding the estimated records due to lack of information sent by the Local Police to Istat for statistical purpose.

The deduplication rule is then applied and it is:

(HSim[DEATHS]>=0.9) or (HSim[INJURED]>=0.9) or (HSim[DEATHS&INJURED]>=0.9)

HSim, the similarity string metric based on the Hamming distance, in the latest version of RELAIS is called 'SimHash' too (§ 3).

Referring to the 2023 data processed, as a concrete example, in the first step, 82,974 records have been considered and submitted to deduplication, and resulting in 353 pairs of possible duplicates to be checked. Of these, 17 pairs were found to be actual duplicates, thus excluding 17 rows that had passed previous checks. Subsequently, an additional set of records (80,343) has been appended to the previous group of records cleaned from the first deduplication, composing the file with 163,317 records for the 2023 survey.

Repeating the deduplication procedure on this new dataset, the process identified 804 possible duplicate pairs, and after further checks, other 19 rows were eliminated as they were confirmed to be actual duplicates. To eliminate the real duplicates, all possible duplicate pairs are "visually" checked, by scrolling through the output created by RELAIS in Excel format. Specifically, the municipality, the month and day of the accident and the plate number of the vehicles involved are at the first stage compared, to avoid the possible mismatching when the information, in different records is reported with swapped position among the three vehicles. The "visual inspection", as a check and coherence step, is crucial to verify the possible match of the first name and surname of the individuals involved. Sometimes, incorrect outcome indications can also be found (for example a person indicated as injured and in a second record classified as deceased, at last identified as the correct outcome).

If the variables mentioned before match perfectly, also the location and other available variables are further checked. Only after all controls of the perfect content coincidence, the record is indicated as a duplicate and eliminated from the survey dataset. In fact, the high number of false positives (pairs identified as possible duplicates and then not eliminated as real duplicates) is not due to a failure of the method but to the subsequent check of other variables (date, time, province and municipality, address and license plates of the vehicles), which reveal that on accident is not clearly doubled, but has to be evidently included without the risk of replication of the same accident, but only base on homonymy.

At last, to summarise, the duplicate record type are classified in two groups: either the incorrect completion of the first accident record by the detecting organisms or the erroneous sending to Istat by different Police corps of two records, referred to the same accident, but containing slight differences.

These processes ensure that the final dataset, without records double counting, is accurate, comprehensive and suitable for detailed analysis and public reporting. This process contribute to a safer road traffic management and policy-making.

References

Cibella N., Fortini M., Mancini L., Scannapieco M., Tosco L., Tuoto T., Valentino L. (2015). *RELAIS User' Guide Version 3.1*. Roma, Istat.

Charikar M.S. (2002). Similarity estimation techniques from rounding algorithms in *Proceedings ACM STOC '02*. Montreal, Quebec, Canada. pp. 380–388

Enamorado T., Fifield B., IMai K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*. 113(2):353-371.

Mancini L., Valentino L., Borrelli F., Marcone L. (2012). Record linkage between large dataset: Evidence from the 15th Italian population census. MMLV 2012. *Quaderni di Statistica* volume 14.

Tuoto T., Bruzzone S., Valentino L., Baldassarre G., Cibella N., Pappagallo M. (2012). Towards an integrated surveillance system of road accidents in *Atti della XLVI Riunione Scientifica (Roma, 20-22 giugno 2012)*, ISBN 978-88-6129-882-8 Editore: CLEUP – Padova

doi.org/10.26398/asaproc.0088

Advancing spatial scenarios adopting generative pre-trained transformers and text-to-image models

Yuri Calleo[a], Francesco Pilla[a], Simone Di Zio[c]
[a] School of Architecture, Planning and Environmental Policy, University College Dublin, Dublin, Ireland.
[b] Department of Legal and Social Sciences, University "G. d'Annunzio", Chieti-Pescara, Pescara, Italy.

1. Introduction

Spatial decision-making (SDM) is defined as the process of making choices considering spatial relationships, geographic data, and the spatial distribution of resources. It particularly involves understanding and considering spatial contexts, including geographic locations, distances, and relationships between variables. Since the SDM involves unpredictable patterns and interactions within a defined physical area (Papadimitriou, 2020), we must consider various elements such as objects, organisms, and processes, as well as their spatial distribution. It therefore follows that spatial complexity is influenced by numerous variables including natural and social aspects, essential to consider for the exploration of potential dystopian scenarios and to better guide current actions in the present (Kohler et al. 2017). In the futures studies (FS) context, spatial scenarios are often developed to depict potential spatial patterns under complex conditions, adopting quantitative models, simulations, or analytical tools to "predict" (*at least from an ideal perspective*), or anticipate dynamic changes over time (Calleo et al. 2023). Nevertheless, since the future is unpredictable, even sophisticated statistical models may not forecast long-term outcomes. Indeed, in FS, there is a shift from traditional forecasting methods to strategic foresight approaches to explore multiple futures and to anticipate, propose, and act with changing circumstances (Martin, 1995).

Among the many methods adopted for the development of spatial scenarios, the Real-Time Spatial Delphi (RTSD) (Di Zio et al. 2017) is identified as a mixed-method approach derived from the traditional Delphi technique and revised for complex spatial contexts. Currently, the RTSD operates in a virtual spatial environment (Calleo and Pilla, 2024), where experts collaboratively respond to different questions by placing one or more spatial points on a dynamic map in real time.

In this system, real-time collaboration and spatial consensus among experts are facilitated. This consensus is measured using a circle C, representing the interquartile range (IQR), which is the most common indicator used in the Delphi technique Generally, experts cooperate to achieve a spatial consensus by identifying an area within a specific territory. The process produces the following outputs: (1) *Spatial outcomes*: geographical maps and spatial analysis results; (2) *Statistical indicator outcomes*: metrics of consensus, indicator outputs, and other relevant non-spatial indicators.

One of the main issues with the RTSD revolves around the communication and dissemination of the results, where a non-expert audience (e.g., local authorities or citizens) might not be aware of the significance of future scenarios. Furthermore, according to scientific literature, RTSD lacks a specific methodology for selecting policies consistent with the SDM process. To address this challenge, we propose a novel hybrid method that combines RTSD with artificial intelligence. This method includes the use of text-to-image models (T2I) for generating visual representations – through images – of potential futures (*Visioning phase*) and generative pre-trained transformers (GPT) for formulating a draft list of potential policies (*Planning phase*) based on experts' judgments derived during the process. This integration aims to streamline the development of spatial scenarios, making them more accessible and actionable for immediate implementation by policymakers.

In this paper, we present a case study focused on climate change threats in Cork, Galway, and

Limerick, Ireland, projected for the year 2050.

2. Method foundations

The novel hybrid method begins with an evaluation of spatial boundaries and climate risks, followed by comprehensive desk research to gather relevant literature and conduct spatial analysis. The method is validated through a case study in Ireland, engaging experts in the cities of Cork, Galway, and Limerick to address diverse regional challenges and inform specific climate policies at local and national levels for the year 2050 (aligning with UN and EU climate change studies).

The desk research starts with a careful review of the scientific literature, EU reports, stakeholder consultations, and relevant data to identify vulnerabilities in the previously defined cities. From what emerged, the three cities are facing different challenges, including flooding, snowfalls, windstorms, and rising sea levels, which can impact both the environment and the population. These drivers are important for formulating the research questions to pose to our panel of experts in the virtual environment. The system adopted in this study is RT-GSCS, a web-based open platform that enables real-time administration of spatial questionnaires adopting the RTSD logic (Calleo et al. 2023).

Once we established the study framework and gathered requisite literature, spatial data, and expert guidance, we proceed by compiling a preliminary list of key environmental factors (K), considered as possible threats to the cities under investigation. These factors are implicitly assessed and documented in municipal-endorsed territorial plans, readily accessible through updated online reports. For each city, we prioritise the top three threats, resulting in $K = 9$ factors. Consequentially, we formulate $RQ = 9$ research questions to pose to our panel of experts, based on their semantic consistency and clarity and attached with x, y coordinates on the system. In Table 1 an overview of the proposed questions is illustrated.

Table 1. Proposed questions.

Area	Research Questions		
	RQ_1	RQ_2	RQ_3
Cork	Which area will be most at risk of flooding?	Which area will be most at risk of impact from snowfall?	Which area will be most at risk of impact from windstorm?
Galway	Which area will be most at risk of flooding?	Which area will be most at risk of impact from snowfall?	Which area will be most at risk of impact from windstorm?
Limerick	Which area will be most at risk of flooding?	Which area will be most at risk of impact from sea levels rise?	Which area will be most at risk of impact from rainfall?

Once the questions are effectively uploaded into the system, we proceed to identify a panel of experts. Invitations are sent via email containing comprehensive technical details, deadlines, guidelines, and system instructions. The panel comprises 176 invited experts, with $E = 79$ participating experts (who contributes at least with a placement of n_i spatial point). The panel includes academic experts, local authorities, industry members, environmental association members, and experts in strategic foresight. Among the participants, 65 are female with the majority affiliated with or from the reference cities, creating a diverse and comprehensive panel for our research objectives. The survey begins by using the experts' emails registered in the system to grant access to the private session link provided by the researchers. Upon accessing the system, experts encounter a real-time spatial questionnaire featuring a dynamic map and sidebar tools. The sidebar prioritizes essential elements, including municipality assessments, prior studies, reference spatial analyses, and survey guidelines. Experts can proceed to address questions based on their expertise, and once a question is selected, they can place one or more n_i point on the map. All input is open to real-time anonymous commentary, fostering dynamic discussion. Following the logic of Di Zio et al. (2017), as a point n_i is placed on the dynamic map, experts observe a dynamic circle C_i that adjusts in size and position based on other points (the adopted algorithm is extensively illustrated in Calleo and Pilla, 2024). Similar to the Delphi logic, if a point n_i is placed outside the circle C_i, it

requires a textual justification, with comments highlighted by a red flag to signal divergence and stimulate consensus. In our case, since the process is in real-time, achieving consensus cannot be considered the only stopping criterion and stability is also important. To address this, we allocate experts a 30-day working period, with the first 20 days dedicated to active participation and the final 10 days focused on achieving stability. During this final step, experts revisit the platform, carefully analyse response patterns, and either refine existing consensus areas or offer justifications for points outside those areas. Then we proceed to implement artificial intelligence models into the process in order to visualize decisions in "reality" and formulate a draft list of policies.

To have a visualisation through images of the depicted spatial scenarios (Sc.), we consider the final circle C derived from the process for each RQ, resulting in a minimum area $min(\boldsymbol{A})$, where $\boldsymbol{A} = [A_1, A_2, \dots, A_N]$ is the vector of the areas of the sequence of consensus circles. The emerged areas depict the specific location considered by experts where policy attention must be focused on the threats considered. Since we have a distribution of $N = n_1, n_2, n_3, \dots, n_N$ points with attached x, y coordinates, to validate our method we extract the coordinates of the n_i point at the centre that forms the last circle C. These coordinates represent one area of the territory and are used to extract via Google Street View, the real image of the area (e.g., a road). Once for each question an image I is extracted, we import it into the Adobe Firefly Image 2 (www.firefly.adobe.com), adopting the "Generative Fill" tool and we provide a prompt to generate the novel image depicting the plausible scenario. In this case, to minimise potential biases arising from the choice of the prompt, we keep the prompts general and use the following prompts: for Cork (Generate… a flooded area (RQ_1), a snow affected area(RQ_2), a windstorm-affected area (RQ_3), for Galway (Generate… a flooded area (RQ_1), a snow-affected area (RQ_2), a windstorm-affected area (RQ_3). Finally, for Limerick (Generate…a flooded area (RQ_1), an area impacted by rising sea levels (RQ_2), a rainfall-affected area (RQ_3). Adopting this approach, the model generates three distinct future visual scenarios for each question, ready for communication and dissemination. Nevertheless, in FS, the development of scenarios should emphasise active policy crafting with the final aim of mitigating future threats and implementing present strategies actually not present in the literature of the RTSD.

To overcome this challenge, we consider the use of GPT-3 (www.chat.openai.com) to generate a preliminary list of policies derived from scenarios identified in the RTSD outputs. The prompt provided to GPT-3 is structured based on the outputs received from the process, in particular: (1) *Key factors*: identifies the specific threat under consideration; (2) *Coordinates*: provides information on the area identified by the last $n_{i(x,y)}$ point resulting in $min(A)$; (3) *Comments*: incorporates all expert feedback, suggestions, and personal opinions in order to provide a specific perspective to GPT-3. The description of the prompt plays a crucial role in shaping the outcomes generated by GPT-3 and for this reason, we adopt the following prompt: "*Act as a policymaker…please generate a list of 10 policies aimed at mitigating the impacts projected for 2050 based on the following information… (1), (2), (3)*. It is important to clarify that these policies serve merely as a foundation for discussion among policymakers and should not be regarded as the final policies to be implemented.

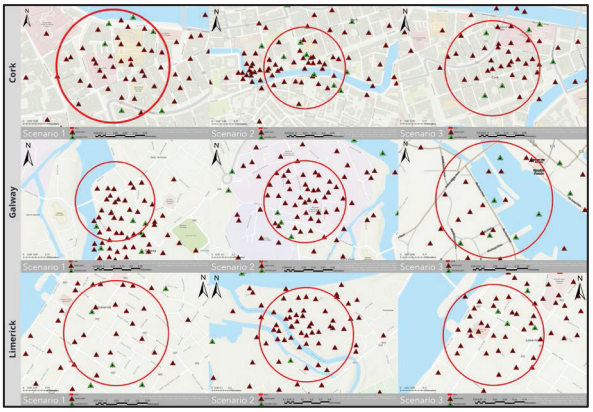
This approach addresses a significant gap in the scientific literature for the final stages of the RTSD process where results are presented, and policies are formulated. With this method, even individuals without specialised expertise can gain insights into potential future scenarios, fostering broader awareness. Moreover, it streamlines the time and resources typically required for workshops or iterative rounds to generate a draft list of territorial policies. This structured approach enhances efficiency and accessibility in the policy formulation process, ensuring a robust response to projected environmental challenges.

3. Method validation

The panel session started on February 1, 2024, and ended on March 2, 2024, spanning a total of $T = 30$ days, during which experts achieved a substantial level of consensus and addressed all

queries in the system with more than $N = 604$ points and 96 comments. Spatial analyses are depicted in Figure 1 and represent the collaborative results derived from experts' contributions after the process. For Cork, experts identified three potential scenarios highlighting climate threats by 2050. These scenarios include a flood risk near the River Lee (Sc.1), vulnerabilities from sudden snowstorms and windstorms in the city centre (Sc.2 and Sc.3), and potential endangerment of vulnerable populations through structural damage, hazards from falling trees, and impacts on transportation and public safety.

Figure 1. Spatial results derived from the process



For Galway, experts identified three scenarios, including a vulnerability to flooding events near the River Corrib (Sc.1), risks from sudden snowfalls in the city centre (Sc.2), and concerns related to intense windstorms in the southeastern area of Galway (Sc.3). Finally, for Limerick, experts expressed concerns about flooding risks near the River Shannon (Sc.1), the repercussions of rising sea levels at the county entry point, and the potential impacts of sudden rain events in the city centre. Overall, the spatial information provided by experts highlights possible climate threats by 2050 that are ready to be assessed by policymakers and governmental bodies.

Geographic maps are not the only output available; we can now present the consensus metrics adopted to deeper focus on the convergence process. In this case, we considered three main indicators: $M_1(FC km^2)$, where FC represents the final circle's area; $M_2 = 1 - FC/S$, calculated as the ratio between the final circle's area (FC) and the city's surface area (S). In this case, the more $M_2 \to 1$ the smaller C, compared to S, thus defining a spatial consensus; $M_3 = 100(FC/IC)$ measures the degree of consensus among experts and is expressed as a percentage, by calculating the ratio between the final circle area FC and the initial circle area IC. For this indicator, the higher the value ($M_3 \to 100\%$) the lower the convergence of opinions among panellists, the closer the value to zero, the higher the achieved consensus. From Table 2, it is possible to depict a strong consensus based on a reduction for all three scenarios of IC compared to M_1, also $M_2 \to 1$ indicating a spatial consensus achieved and the values of M_3 are very low (we can considered achieved a consensus when $M_3 \leq 20\%$).

Table 2. Measures of spatial consensus

City	Scenario	$S\ (km^2)$	IC	$FC\ (M_1)$	M_2	$M_3(\%)$	N	M
Cork	Sc.1	187	1.41	0.16	0.999	11%	71	10
Cork	Sc.2	187	2.28	0.18	0.999	7%	88	13
Cork	Sc.3	187	1.28	0.14	0.999	10%	58	9
Galway	Sc.1	57.3	1.32	0.17	0.997	12%	80	8
Galway	Sc.2	57.3	2.42	0.10	0.998	4%	90	12
Galway	Sc.3	57.3	1.51	0.16	0.997	10%	58	9
Limerick	Sc.1	61.3	1.41	0.20	0.996	14%	44	6
Limerick	Sc.2	61.3	3.10	0.32	0.994	10%	60	8
Limerick	Sc.3	61.3	2.55	0.14	0.997	5%	55	11
Total							**604**	**96**

Following the presentation of the primary outputs from the previous analysis, we employed T2I models to visualise potential future scenarios, with an average generation time of $\bar{x} = 3$ seconds. Figure 2 illustrates nine distinct scenarios for each city, incorporating authentic territorial images augmented with AI-generated elements. Sc.1 across all three cities effectively simulates potential flooding events, offering a realistic representation of anticipated conditions for the year 2050. These visualizations avoid idealised projections, thereby providing a clear and accessible depiction of the imminent risks associated with the actual imagery. In terms of atmospheric impacts, the models depict several future challenges: Sc. 2 for Cork and Galway portrays potential snowfall, Sc. 3 for Cork and Galway illustrates potential windstorms, and Scenario 3 for Limerick demonstrates the effects of heavy rainfall. Additionally, Sc. 2 for Limerick visualizes the projected rise in sea levels leading to potential flooding near riverbanks. These visual representations serve to effectively convey the potential environmental challenges that the cities and their residents may face in the future.

Figure 2. Visual scenarios using text-to-image models



This implementation marks a significant innovation within the RTSD method, extending its application beyond territorial planning to diverse domains such as product development, corporate strategies, and security planning. Traditional graphic design software often suffers from biases introduced by human designers and inefficiencies in resource use. In contrast, AI models – though still in development and with their own limitations – hold significant potential for improving decision-making processes. After the visioning phase, we integrate the results into GPT-3 to generate a draft list of potential policies for refinement and consultation with policymakers and local authorities. An example of the developed policies is illustrated in Figure 3 (for more analytic results, please refer to this link: https://tinyurl.com/nrjx47ap). Based on the results, to address flood risk, the strategies focus on several key approaches: investing in flood defences, incorporating green solutions, engaging the community, and implementing climate-conscious planning to strengthen urban resilience.

Similarly, to prepare for potential snowstorms, the policies emphasize winter preparedness, infrastructure resilience, and community involvement to ensure public safety and service continuity. For windstorm vulnerabilities, the strategies include enforcing resilient building codes, enhancing emergency readiness, investing in green infrastructure, and fostering intersectoral collaboration to reduce damage and build community resilience. In summary, Limerick addresses both flood and windstorm risks through a comprehensive strategy that includes improving infrastructure, boosting community resilience, and managing natural resources. These policies represent proactive efforts to mitigate climate risks and ensure the city's long-term viability in the face of evolving environmental challenges.

Figure 3. An example of developed policies with GPT-3



3. Concluding remarks and recommendations

In this study, we introduced an innovative enhancement to the RTSD method by incorporating advanced AI models to improve spatial scenario communication and policy development. Our approach adopts T2I models and ChatGPT to bridge spatial and statistical components, offering a comprehensive understanding and visualization of future scenarios. This integration addresses a critical gap in scenario planning by providing realistic visual depictions of potential spatial arrangements and generating preliminary lists of territorial policies for evaluation. Our methodology not only streamlines policy development but also alleviates the burden on policymakers by offering actionable insights and visual representations of complex spatial challenges. By enhancing the RTSD method with T2I models, we enable stakeholders to better understand and visualize future scenarios, facilitating more informed decision-making. Additionally, ChatGPT aids in drafting territorial policies, accelerating the policy development process, and supporting effective decision-making.

Our contribution lies in enriching the RTSD framework with advanced AI tools that provide clearer scenario depictions and actionable policy drafts, thus making future challenges more tangible and relevant for discussion. This advancement supports long-term engagement in scenario development and aligns with the need for clear, actionable insights into policy operations. Future research will explore refining AI models, validating their effectiveness, and assessing long-term impacts to further advance AI integration in spatial scenario planning and develop more sustainable, resilient spatial strategies. Additionally, we are currently working on transforming the spatial scenarios into short videos, which possess a greater communicative power than static images.

References

Calleo, Y., Di Zio, S., Pilla, F. (2023). Facilitating spatial consensus in complex future scenarios through real-time spatial Delphi: A novel web-based open platform. *Futures & Foresight Science*, **5**(3-4).

Calleo, Y., Pilla, F. (2024). Optimizing spatial survey administration adopting RT-GSCS: A statistical perspective on performance metrics. *MethodsX*, **12**, 102578.

Di Zio, S., Rosas, J.D.C., Lamelza, L. (2017). Real time spatial Delphi: Fast convergence of experts' opinions on the territory. *Technological Forecasting and Social Change*, **115**, pp. 143-154.

Kohler, M., Stotten, R., Steinbacher, M., Leitinger, G., Tasser, E., Schirpke, U., Schermer, M. (2017). Participative spatial scenario analysis for alpine ecosystems. *Environmental Management*, **60**, pp. 679-692.

Martin, B.R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management*, **7**(2), pp. 139-168.

Papadimitriou, F. (2020). *Spatial Complexity: Theory, Mathematical Methods and Applications*. Springer.

doi.org/10.26398/asaproc.0089

Enhancing mobility and tourism through data analytics and generative AI

Denis Cappellari , Glauco Mantegari , Bruno Zamengo,
Motion Analytica, Udine, Italy.

1. Introduction

Today, vast amounts of data from mobile operators, floating car data (FCD), traffic loops, public transport ticketing, and GPS apps generate valuable insights for decision-makers in urban planning, tourism, and sustainable mobility.

This presentation showcases applications where anonymized big data - telecommunications, FCD, and GPS - complement official statistics to analyze urban presence and mobility. These insights help quantify visitor patterns, transport preferences, traffic congestion, trip lengths, and more. In urban planning, mobility data reveal commuting patterns, peak traffic times, and public transport usage, aiding infrastructure optimization. In tourism, visitor flow analysis improves resource allocation and enhances experiences. In sustainable mobility, data-driven insights help reduce congestion and promote eco-friendly transport.

Finally, advancements in large language models (LLMs), such as OpenAI's GPT, Google's Gemini, and Meta's Llama, enable intuitive data analysis beyond SQL queries and dashboards. We will explore how these technologies make insights accessible to a broader audience, including non-technical users.

The presentation is organized as follows: Section 2 is dedicated to tourism, Section 3 to the analysis of civil aviation, Section 4 to the study of mobility, and Section 5 to our approach to using AI and Generative AI.

2. Tourism

We first explored mobile big data in tourism to provide faster, more detailed statistics than official sources. Telecom data has been widely studied for assessing tourism's impact on residents (Baldin et al., 2024), analyzing overtourism (Bertocchi et al., 2021), and enhancing official statistics, F. Ricciato, et al., (2018), E. Saluveer, et al., (2020), United Nations Committee of Experts on Big Data and Data Science for Official Statistics (2021), Nurmi et al., (2019), Ricciato, et al., (2021).

A study published in 2024, (Scotti et al., 2024), based on data provided by Motion Analytica and Vodafone Business Italia, addresses the drivers of overnight stays and same-day visits in the tourism of the Lombardy region of Italy. Mobile network data proved to be more granular and richer than official statistics and not affected by the biases that are typical of other sources (e.g., geolocated photos), which are more commonly used by specific segments of the population.

Key considerations when using mobile network data include its uneven infrastructure distribution, favoring dense areas and affecting precision by region. To address this, we apply techniques ensuring robust statistics, with greater accuracy in well-covered areas and larger-scale phenomena. Additionally, as the data reflects only smartphone users, we account for this limitation when extrapolating results, integrating other sources and official statistics where possible.

The produced statistics were then used to study which factors might influence the duration of stays, seeking correlations with the presence of specific services (e.g., ski slopes) or specific points of interest (e.g., cultural attractions).

Thanks to the ability to study the behaviour of day-trippers — considered one of the additional strengths of this data source — the authors were able to verify that events such as festivals or specific

intermodal nodes influence the level of day-trippers attracted to a destination.

The authors also suggest that this type of data and the derived analyses can be useful for defining specific policies for a territory or a group of nearby destinations. They propose, for example, the identification and creation of common hubs for accommodation, transport, and major attractions, such as cultural sites, using more peripheral areas to offer various leisure and recreational attractions that will draw in day-trippers.

3. Aviation

The world of civil aviation includes two main types of actors: airports and airlines. For an airport, a strategic aspect is its centrality within the network of air connections to attract as many travelers as possible.

The centrality of an airport can be viewed from two perspectives: inbound traffic analysis, which refers to the airport's capacity to serve as an entry point for travelers wishing to reach the airport's region, and outbound traffic analysis, which refers to the airport's capacity to act as a gateway to remote destinations of interest for people living near the airport.

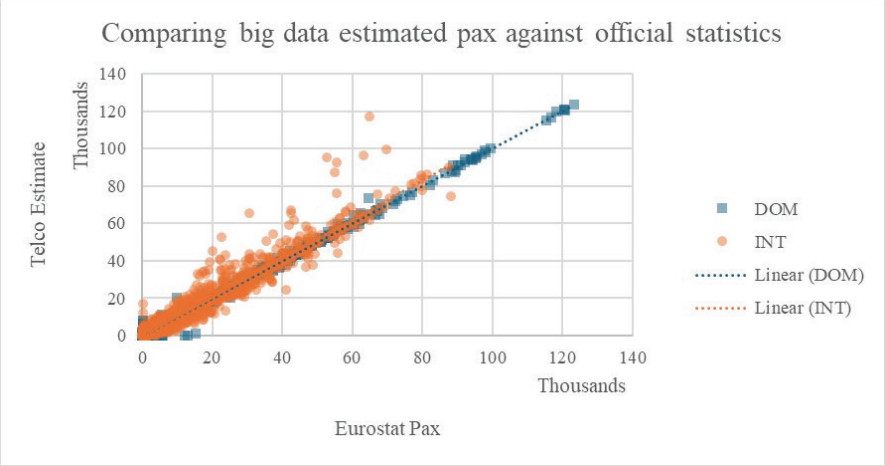
In this context, mobile phone data, cross-referenced and enriched with other data sources such as timetables, has proven to be a valuable resource (Motion Analytica, 2022) compared to other market solutions or official sources, (Associazione Italiana Gestori Aeroporti, 2022), (Eurostat, 2022). This data provides airports with tools to estimate their catchment and distribution areas based on the actual behavior of users.

These basic models were initially supplemented by an unsupervised study using clustering techniques, which gradually evolved into more sophisticated machine-learning solutions that integrated mobile phone data with the actual departure/arrival times of flights at airports. The integration of official statistics from a preceding period allowed for precise expansion of estimates from the sample population to the entire target population including all individuals who flew from any of the over forty major Italian airports, not just Vodafone Italy customers or those roaming on their network. In Figure 1 we report the comparison of our estimates with official data for routes (see Eurostat, (2022), month of August 2022) showing that the overall correlation per route exceeds 99% for domestic flights and 95% for international flights, with linear coefficients of 1.00 and 1.04, respectively. In Figure 2 additionally, as expected, we inspected the MAE index (see Morley *et al.*, 2018), which indicates that the accuracy of the estimates increases with the size of the phenomenon: routes with higher passenger traffic report significantly lower MAE indices. Consequently, where phenomena are smaller and variability is greater, we typically aggregate these data into statistics that describe larger phenomena, without adding additional heuristics or forcing the data.

One of the added values of big data estimates compared to official statistics is the ability to investigate additional details. In this case, for example, it is possible to delve into route data concerning the estimated habitual residence of users based on their behavior.

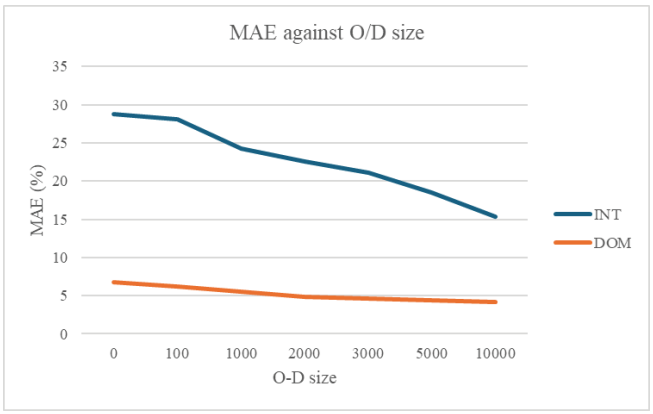
The maps in Figure 3 show the market share among the airports of Ancona, Bologna, Firenze, and Pisa by the municipality of residence of travelers. The first map covers all international routes while the other map shows routes to Germany. For privacy reasons, only municipalities with a certain minimum number of travelers are reported.

The ability to determine the origin of travellers from a particular airport or even a specific route can be used to understand the airport's catchment area based on actual user behavior. It can also help quantify phenomena such as leakage, where travellers choose to use an airport that is farther from their home because they might be interested in a route not available at the nearby airport, or due to the convenience of the airline's offerings, or better connections at the farther airport. For instance, figures extracted from image Figure 3 show that Bologna airport appears to be the strongest, having the broadest catchment area. Florence and Pisa show an interesting pattern: overall, Pisa's catchment area is larger than Florence's but, when analyzing connections with Germany, Florence's catchment area becomes more extensive.



Market	R^2	Slope
DOM	99%	1.00
INT	95%	1.04

Figure 1 The x-axis shows number of passengers recorded by official statistics, while the y-axis our big data estimated value.



O-D size	INT	DOM
0	28.74	6.73
100	28.09	6.22
1000	24.27	5.51
2000	22.61	4.85
3000	21.11	4.67
5000	18.53	4.44
10000	15.31	4.15

Figure 2 We grouped O-D (origin-destination) pairs into bins depending on monthly passenger count. The "O-D size label" represents the bin's lower bound (e.g. "100" indicates O-D pairs with 100-999 passengers). The blue line and the "INT" column in the table show the MAE index for international O-Ds, while the orange line and the "DOM" column refer to domestic O-Ds.

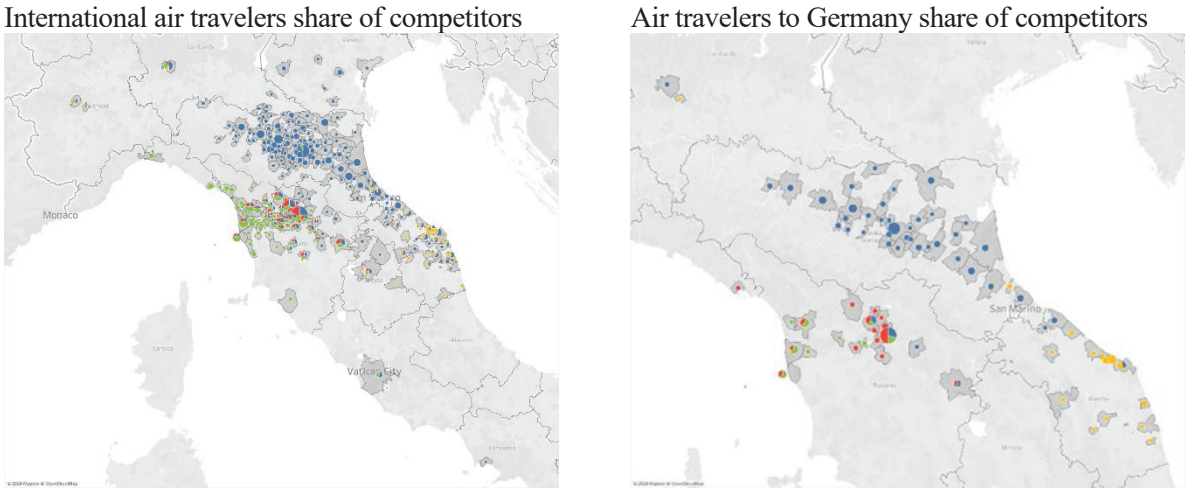


Figure 3 Air traveller market share distribution among Ancona (yellow), Bologna (blue), Firenze (red) and Pisa (green) airport at municipal level. Each pie chart represents the share of travellers in various municipality. The maps provide a visual comparison of competitor dominance and regional preferences in each travel category.

4. Big data for general mobility

The study of mobility through big data and ML/AI experienced significant acceleration during the COVID-19 pandemic, when numerous studies were developed using different data sources to detect changes continuously and extensively in people's movement habits.

Some studies focused on the use of GPS data, as seen in (DePhillipo et. al., (2021), where Google COVID-19 mobility reports were used alongside with social distancing measures to find significant associations between geographic activity and prevalence of COVID-19 infections in the U.S Other studies leveraged mobile network data to monitor mobility behaviors during the pandemic (Jewell et al., 2021), (Xiong et al., 2020), (Calabrese et al., 2021). These studies represent an evolution of earlier research (Calabrese et al., 2011), (Ricciato et al., 2020) which explored the possibility of using mobile network data to measure mobility. For instance, they developed Origin-Destination matrices to study actual mobility demand on a broad sample. Moreover, GPS data tend to be biased since they are generated by a specific population that uses this technology more frequently. Consequently, a strategy to obtain more robust statistics consists of combining insights derived from both sources—MNO and GPS—leveraging their respective strengths.

In addition to these data sources, which typically detect the movement of people, other contexts have seen the development of studies examining the mobility of specific categories of vehicles. For example, the use of floating car data (FCD) from vehicle black boxes was explored in (Mannini *et al.*, 2018) and (Nalubega, 2023) for examining parking demand, estimate parking search time, and measure traffic congestion. Finally, in (Costa & Silvestri, 2021) and (Cavattoni et al., 2024) shared mobility data, such as bicycles and scooters, was used to optimize bike spreading and to measure COVID-19 impact on shared mobility. In the case of FCD and shared mobility data, these are closely tied to the market penetration of the provider and, since they pertain to vehicles, are not directly connectable to other sources focused on people. However, the integration of various sources, while accounting for their limitations and unique characteristics, can provide a more nuanced and complete picture of the phenomenon.

As in other sectors, the use of big data for mobility and the insights derived from it can serve various purposes. Two major Italian institutions, (Ministero delle Infrastrutture e della Mobilità Sostenibili, 2022) and (Tartaglia *et al.*, 2024), identify the following main applications: (i) Measure the real demand for mobility comprehensively, detailing it by means of transportation; (ii) Plan and improve public transport services and infrastructure; (iii) Plan the ecological transition to support decisions that promote the shift towards more sustainable mobility and (iv) Monitor the evolution of mobility behaviors over time and measure the effectiveness of chosen strategies and implemented interventions.

These data have also been used to explore the possibility of supplementing official statistics. For example, in our joint work with ISTAT and Vodafone (Radini et al., 2021), we studied the population that habitually resides in an area and constructed commuting matrices. From these experiments and other experiences by official statistical agencies, several European projects have emerged with the aim of defining a common framework for all member countries. This framework seeks to supplement traditional sources with these new data sources in the production of official statistics for the European Community (Ricciato et al., 2018), (Ricciato, 2018), (Ricciato et al., 2021).

5. AI & GenAI

In November 2022, ChatGPT was announced and released to the public, creating a new wave of interest in AI and ML. However, several works utilizing these types of technologies had already been in use for some time. For example, the solution presented in Section 3 also employs typical ML techniques to correctly associate a passenger with a flight.

A publicly documented case, which is more extensively covered, can be found in (Kim *et al*.

(2022) and further explored in its bibliography), where the authors propose a machine learning model to classify modes of travel based on movement detected by GPS data, using a convolutional neural network.

Thanks to recent technological advancements pioneered by OpenAI, followed by contributions from Google, Microsoft, and Meta, Large Language Models (LLMs) are proving to be an extremely versatile tool. One potential application we explored in (Padoan *et al.*, 2024), involves the use of the reasoning and communication abilities of a LLM to interact with data through natural language. This approach reduces the dependence on dashboards or SQL queries, thereby making data more accessible to a broader audience. To implement this, we designed a cloud-based architecture using (i) Langchain, a Python library designed to streamline the creation of conversational AI applications; (ii) Llamaindex, a data framework designed to support the development of applications based on LLMs; (iii) Azure AI, cloud-based service offering a variety of AI tools for building and deploying AI applications and (iv) Chainlit, an open-source Python package designed for building production-ready conversational AI applications.

From our experiments, we found that by appropriately adapting the general model to specific datasets and providing the LLM with the necessary context, it is possible to create chatbots capable of answering domain questions based on both structured and unstructured data. This required instructing the application with all the contextual elements necessary to interpret the data structure and the semantic value of the contained data. This was achieved through a JSON file containing these metadata and several examples of questions and answers shown to the chatbot engine. The chatbot developed for these experiments was able to deliver basic descriptive statistics as well as perform more complex analyses that required additional computations. Moreover, by equipping the LLM with the right tools, we were able to use it to generate customized charts.

References

Associazione Italiana Gestori Aeroporti (2022). Tabella dati di traffico.

Baldin, A., Bertocchi, D., Camatti, N. (2024). Crowding-in and (temporary) crowding-out in Venice. The effect of cultural events on residents. *Tourism Economics*, 0(0).

Bertocchi D., Camatti, N., Van der Borg, J. (2021) Tourism peaks on the three peaks. Using big data to monitor where, when and how many visitors impact the Dolomites UNESCO World Heritage Site, *Rivista Geografica Italiana*, vol. 3.

Calabrese, F., Ferrari, L., Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys*, 47(2), pp. 1–20.

Calabrese, F., *et al.* (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4), pp. 36–44.

Calabrese, F., *et al.* (2021). Using Vodafone mobile phone network data to provide insights into citizens' mobility in Italy during the Coronavirus outbreak. *Data & Policy*, 3, p. e22.

Cavallo, L., *et al.* (2022). Exploring mobile network data for tourism statistics: The collaboration between Istat and Vodafone Business Italia. *Rivista di Statistica Ufficiale*, 3, pp. 43–76.

Cavattoni, M, Comin, M., Silvestri, F. (2024). COVID-19 pandemic's enduring impact on urban mobility: The case of free-floating bike sharing in Padova, Italy, *Transportation Research Record.*

Confente, I., *et al.* (2024). Integrating tourists' walk and talk: a methodological approach for tracking and analysing tourists' real behaviours for more sustainable destinations. *Journal of Sustainable Tourism*, 1, pp. 1–21.

Costa, E., Silvestri, F. (2021). On the bike spreading problem.

DePhillipo, N. N. et al., (2021) Mobile phone GPS data and prevalence of COVID-19 infections: Quantifying parameters of social distancing in the U.S.,*Arch. Bone Jt. Surg.*, 9, (2), pp. 217–223,

Eurostat (2022). Air transport measurement – passengers and freight.

Jewell, S., et al. (2021). It's complicated: Characterizing the time-varying relationship between cell phone mobility and COVID-19 spread in the US. *npj Digital Medicine*, 4(152).

Kim, J., Kim, J. H., Lee, G. (2022). GPS data-based mobility mode inference model using long-term recurrent convolutional networks. *Transportation Research Part C: Emerging Technologies*, 135, p. 103523.

Mannini, L., Cipriani, E., Crisalli, U., Gemma, A. (2018). FCD data for on-street parking search time estimation. *IET Intelligent Transport Systems*, 12, pp. 664–672.

Ministero delle Infrastrutture e della Mobilità Sostenibili (2022). *Mobilità e logistica sostenibili – Analisi e indirizzi strategici per il futuro*.

Motion Analytica (2022). Analysis of the catchment area of Milan airports.

Morley, S. K., Brito, T. V., Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, 16, pp. 69–88.

Nalin, A., *et al.* (2024). Assessing veracity of big data: An in-depth evaluation process from the comparison of mobile phone traces and ground truth data in traffic monitoring. *Journal of Transport Geography*, 118,

Nalubega, S. I. (2023). Development of a floating car data (FCD) model to evaluate traffic congestion: A case of Kampala, Uganda. Stellenbosch University.

Nurmi, O., Piela, P. (2019). The use of mobile phone data in tourism statistics. In *Proceedings of the Special Topic Session*, 62nd ISI World Statistics Congress, vol. 4, pp. 135–142, Kuala Lumpur, Malaysia, Aug. 18–23.

Padoan, L. et al. (2024). Mobility ChatBot: Supporting decision making in mobility data with chatbots. In *Proceedings of the 25th IEEE International Conference on Mobile Data Management (MDM)*, Brussels, Belgium, pp. 295–30.

Radini, R. et al. (2021). Come i dati di telefonia mobile possono contribuire alle statistiche sperimentali sulla mobilità.

Ricciato, F. et al. (2018). Processing of mobile network operator data for official statistics: The case for public-private partnership. Presented at the *DGINS Conference*, *Bucharest, Romania, Oct. 10–11.*

Ricciato, F., (2018). Towards a reference methodological framework for processing MNO data for official statistics, presented at the *15th Global Forum on Tourism Statistics*, *Cusco, Peru, Nov. 30, 2018.*

Ricciato, F. et al. (2020). Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing*, 68.

Ricciato, F. *et al.* (2021). A proof-of-concept solution for the secure private processing of longitudinal Mobile Network Operator data in support of official statistics. *Eurostat and Cybernetica*.

Ricciato, F., Coluccia, A. (2023). On the estimation of spatial density from mobile network operator data. *IEEE Transactions on Mobile Computing*, 22(6), pp. 3541–3557.

Scotti, F. et al. (2024). Exploring drivers of overnight stays and same-day visits in the tourism sector. *Scientific Reports*, 14, p. 9840.

Saluveer, E. et al. (2020). Methodological framework for producing national tourism statistics from mobile positioning data. *Annals of Tourism Research*, 81.

Tartaglia, M. et al. (2024). Methodological aspects for passenger mobility analysis using big data.

United Nations Committee of Experts on Big Data and Data Science for Official Statistics - UN-CEBD. (2021). *Mobile phone data*. United Nations, NY, USA.

Xiong, C. et al. (2020). Mobile device location data reveal human mobility response to state-level stay-at-home orders during the COVID-19 pandemic in the USA.

doi.org/10.26398/asaproc.0090

An innovative approach for the analysis of socio-economic phenomena at sub-municipal level: The Household Deprivation Study project.

Giancarlo Carbonetti, Flavio Biasciucci, Andrea Cutillo,
Matteo Mazziotta, Valeria Quondamstefano, Debora Tronu,
ISTAT, Rome

1. Introduction

Over the past decades, NSIs have shown an increasing focus on the demand for sub-municipal data. In this regard, the Italian National Institute of Statistics (Istat) has undertaken a significant revision of the production processes of official statistics. In particular, the Integrated Register System (IRS) and the Permanent Population and Housing Census (PPHC) should be mentioned.

The development of statistical methodologies and the adoption of advanced IT solutions have made it possible to integrate administrative data with surveys results and to geo-code information at a very fine territorial level. This has made it possible to define an integrated and geo-referenced database for the territory, rich in annually renewable information on the main demographic, social and economic variables of individuals, household types and certain characteristics of dwellings.

The new process is continuously evolving to meet ever higher quality standards, release data with high timeliness, and ensure a constant increase in the supply of information at the highest territorial level. These results offer new perspectives for the analysis of the socio-economic phenomena of the population at sub-municipal level, both on spatial and temporal dimensions.

In this regard, Istat is conducting an experimental project to study the socio-economic deprivation of households at the sub-municipal level. The possibility of identifying the most critical areas within the municipality represents an extraordinary opportunity for policy makers to define intervention programmes to combat household deprivation in a territorially targeted manner. Moreover, the possibility of replicating the analyses over time will also make it possible to assess the impact of the interventions implemented in the territory.

2. The new sub-municipal data production process with the PPHC in Italy

To replace the decennial census, in 2018 Istat launched the Permanent Population and Housing Census (PPHC): a combined approach which integrates administrative data and sample surveys, with yearly data collection and dissemination of data (Falorsi, 2017; Gallo and Zindato, 2018). The development of the new strategy was facilitated by the availability of data from administrative sources and the opportunity to use the statistical registers built in Istat.

The core of the PPHC is the Basic Register of Individuals (BRI) which contains the persons annually identified by the census as usual residents in Italy. Together with the Basic Register of Places (BRP) and the thematic registers on education and employment, the BRI forms the basis for the production of population census data in a combined census design, with two ad hoc sample surveys (area survey and list survey) conducted annually to collect data for census variables and information useful for the population count. Concerning the census outputs, the PPHC produces: fully register-based population count; census hypercubes estimated by the joint use of information already available in registers and of data collected on the field.

In the context of the PPHC, the data production process at the sub-municipal level is based on the linkage of BRI and BRP (Carbonetti *et al.*, 2024b), which links individuals and households with their housing and buildings, and geo-codes all statistical units in an unambiguous and consistent way. This process allows for the dissemination of both census data and administrative data from BRI-linked sources at very detailed spatial levels.

This is an evolving process due to the availability of new and updated data sources and continuous methodological advances in linkage and geo-referencing techniques. The main features of the sub-municipal data production model are: consistency in small domains of data on populations, households, dwellings and buildings; integrated use of spatial data sources for geo-coding; use of administrative data referring to dwellings and buildings; strong focus on quality[1] (geographical accuracy; data validation); possibility to repeat the process every year.

3. Impact on sub-municipal data offer and new scenarios for spatial studies

The availability of data by enumeration areas (EA) derived by PPHC and the new systems of dissemination, will be different overall from that of past censuses. There will be less data than in the traditional census, but also new information thanks to the possibility of using administrative sources and statistical registers, in an integrated way with the BRP and geo-referencing the information to the territory. Furthermore, Istat is enhancing sub-municipal data dissemination in order to make geographical tools more easily accessible to less skilled users. New forms and tools for the dissemination of sub-municipal data are being explored, with the aim of visualising complex topics in a simple way and illustrating territorial and local differences at a glance.

The new process will lead to the availability of more accurate and timely data on an annual basis. This opens up new possibilities for spatial analysis. Linking administrative data to census results makes it possible to construct specific indicators to study demographic, social and economic phenomena on the population as never before. Moreover, the geo-coding of information to EAs and the annual replication of the available information base opens up a new frontier for studying the dynamics of socio-economic population phenomena on the territory. It will in fact be possible:

- follow the changes and make comparisons on both spatial and temporal dimensions;
- identify the sub-municipal areas in which the phenomenon is most concentrated.

4. The project for the study of household deprivation at sub-municipal level

Istat is conducting, together with some municipalities[2], an experimental project to study socio-economic deprivation of population at sub-municipal level (Carbonetti *et al.*, 2024a). The project exploits the richness of information from censuses, administrative sources and statistical registers held by Istat. Once the concept of socio-economic deprivation was defined[3], an exploration of the sources was carried out to identify the most appropriate ones - in terms of data availability and quality - useful to calculate specific individual indicators (referring to households or individuals) at the highest territorial detail (EAs). These sources must have the following characteristics: adequate level of quality (*accurate*); reference to a well-defined time interval (*timely*); renewable from year to year (*updatable*); allow geo-reference to the enumeration areas (*geo-codifiable*).

Based on the availability of census results and administrative data, nine individual indicators covering different components of deprivation (economic, employment, education, and housing) were defined and calculated at the EA level (Table 1). In order to synthesize the individual indicators in a unique measure, the *Socio-Economic Deprivation Index (SED-Index)* at EAs' level, the Adjusted Mazziotta-Pareto Index (AMPI) is used, because the influence analysis demonstrates the validity compared to other methods in terms of robustness. AMPI is a partially non-compensatory composite index based on a standardization of the individual indicators, at the reference time, that makes the indicators independent from the unit of measure (De Muro *et al.*, 2011). Therefore, all the individual indicators are assigned equal weights and absolute time comparisons are allowed (Mazziotta and Pareto, 2016; Mazziotta and Pareto, 2017). In fact, a re-scaling of the individual indicators in the

[1] The quality of data is ensured by the use of the BRP, which enables correct geo-coding of units. This results in a consistency of count data on individuals, households, dwellings and buildings. Prior to dissemination, the data are validated with benchmark information.

[2] Bologna, Florence, Gorizia, Messina, Perugia, Milan, Modena, Palermo, Prato, Olbia, Rome, Taranto, Verona, and Parma.

[3] The definition chosen is as follows: *"A condition in which households and individuals experience difficulties in adequately meeting their basic needs due to a lack of or insufficient economic, employment, educational, social and housing resources and opportunities"*.

range (70; 130) according to two 'goalposts' is proposed, i.e., a minimum and a maximum value which represent the possible range of each variable for all time periods and for all units.

Table 1 – Individual indicators of socio-economic deprivation: an initial proposal.

Individual indicator		Component	Sources(*)
Dep1	Individuals aged 67 and over living alone, without home ownership (%)	Economic	PPHC, Cadastre
Dep2	Households in which no member is employed or receiving a pension (%)	Economic	PPHC, INPS
Dep3	Households not living in owner-occupied dwellings (%)	Economic	PPHC
Dep4	Employees aged 35-64 with a temporary employment relationship (%)	Occupational	PPHC, INAIL, INPS
Dep5	Employment rate 35-64 years	Occupational	PPHC
Dep6	Individuals aged 25-64 without upper secondary education (%)	Educational	PPHC
Dep7	Individuals aged 15-29 not working and not attending any regular course of study (%)	Educational	PPHC, MIM, MUR
Dep8	Individuals aged 18-24 without upper secondary education and not in education (%)	Educational	PPHC, MIM, MUR
Dep9	Ratio of occupants to rooms in occupied dwellings	Housing	PPHC, Cadastre

(*) Sources: PPHC (Istat); National Social Security Institute (INPS); National Institute for Insurance against Accidents at Work (INAIL); Ministry of Education (MIM); Ministry of University and Research (MUR); Register of dwellings and buildings (Cadastre).

A specific procedure is being developed to identify the critical areas of highest concentration of cases of socio-economic deprivation within the municipality. These areas will be determined by the aggregation of contiguous enumeration areas around EAs with the highest SED-Index values, according to statistical (internal homogeneity) and geographical (shape and extent) criteria.

The results will be a useful tool for municipal administrators in planning local policies. Moreover, the possibility of replicating the study each year will make it possible to follow the changes of the phenomenon and assess the effectiveness of the interventions implemented.

5. A case study: socio-economic deprivation in the municipality of Rome

This section presents some results of a preliminary study on the socio-economic deprivation of population in the municipality of Rome. For the purposes of the analysis, only EAs located in urban areas where population and residential buildings are present were considered: from the total number of EAs in Rome defined in 2021 (23,454), the observation was reduced to 11,691. This implies the exclusion of all the EAs of the extra-urban territory and the EAs that, although part of the urban territory, have no population or very few residents (examples: churches, hospitals, barracks, monuments, parks). The first step was to define the database for calculating the individual indicators (Table 1) at EA level. Subsequently, the values of the indicators were synthesised, as mentioned, using the AMPI methodology to determine the SED-Index. Table 2 shows the minimum, median, mean and maximum of the distribution of the calculated values of the individual indicator and the SED-Index relative to the EAs in Rome involved in the study.

Table 2 – Minimum, median, mean and maximum values of the individual indicators and the SED-Index for the EAs of the municipality of Rome involved in the study.

	Dep1	Dep2	Dep3	Dep4	Dep5	Dep6	Dep7	Dep8	Dep9	SED-Index
Minimum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	91.3
Median	7.0	10.3	31.0	0.0	72.4	19.2	19.1	3.3	0.4	100.1
Mean	9.4	11.5	34.7	1.1	70.7	21.4	20.6	8.8	0.4	100.9
Maximum	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	5.0	129.5

Source: Istat (2021).

The validation phase aims at assessing the robustness of the composite index in terms of its ability to produce correct and stable measures and its discriminatory power. Indeed, the results and the classification of statistical units according to a composite index may depend to a large extent on the

choices made in the previous stages (choice of individual indicators, normalisation and aggregation, etc.). For this reason, it is necessary to carry out statistical analyses to examine the robustness of the classifications when an elementary indicator is included/excluded or when the methods used to construct the composite index vary.

In this regard, an influence analysis showed that the AMPI method is the most robust for this type of synthesis. In particular, this analysis shows by how many positions on average the ranking of each territorial unit shifts if one indicator is removed at a time (Figure 1). The most influential indicator is the Dep7, and the least influential is the Dep4.

Figure 1: Influence Analysis of individual indicators

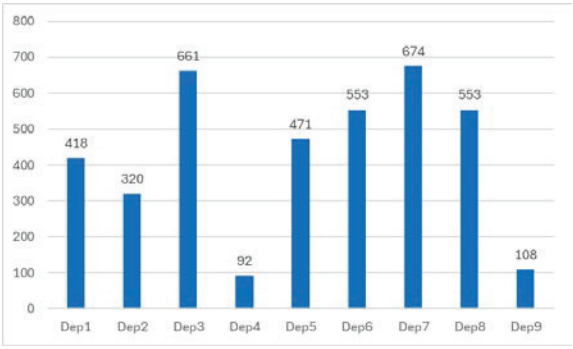


Table 3 shows the distribution of SED-Index values calculated on EAs and "Urban Areas" (UA), one of the administrative subdivisions[4] of the territory of the municipality of Rome.

Table 3 – Distribution of SED-Index calculated for the 11,691 EAs involved in the study, and UAs in the municipality of Rome (five UAs were excluded because without population).

SED-Index	No. of EAs	%	No. of UAs	%
< 95	154	1.3	25	16.6
95 - 100	5,565	47.6	51	34.0
100 - 105	4,528	38.7	34	22.7
105 - 110	1,005	8.6	31	20.7
110 - 115	331	2.9	8	5.3
115 - 120	82	0.7	1	0.7
> 120	26	0.2	-	-
Total	11,691	100.0	150	100.0

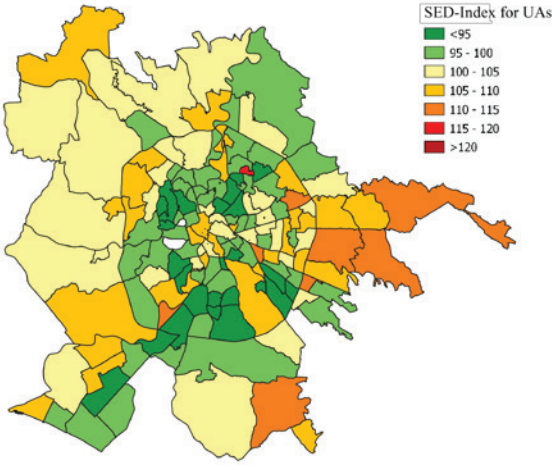
Source: Istat (2021).

From the distribution of the SED-Index values on the EAs, it can be seen that for 48.9% of the cases the value is lower than the municipal data used as a reference base (=100); for these EAs it can be assumed that no situations of deprivation are present. In 3.8% of the cases, on the other hand, very high deprivation situations are observed (SED-Index>110); in particular, in 108 EAs (less than 1%) the SED-Index indicates cases of severe deprivation (SED-Index>115). Turning then to the most critical SED-Index values (>110) in the UAs, these are 6% of cases (slightly less than twice the percentage value observed for EAs in the same class).

Therefore, the increase in territorial detail makes it possible to better identify the real cases of deprivation in the territory under consideration.

Figure 2 represents the spatial distribution of SED-Index values referred to the UAs of Rome. The highest cases of deprivation are observed in the urban areas to the west and east of the municipality.

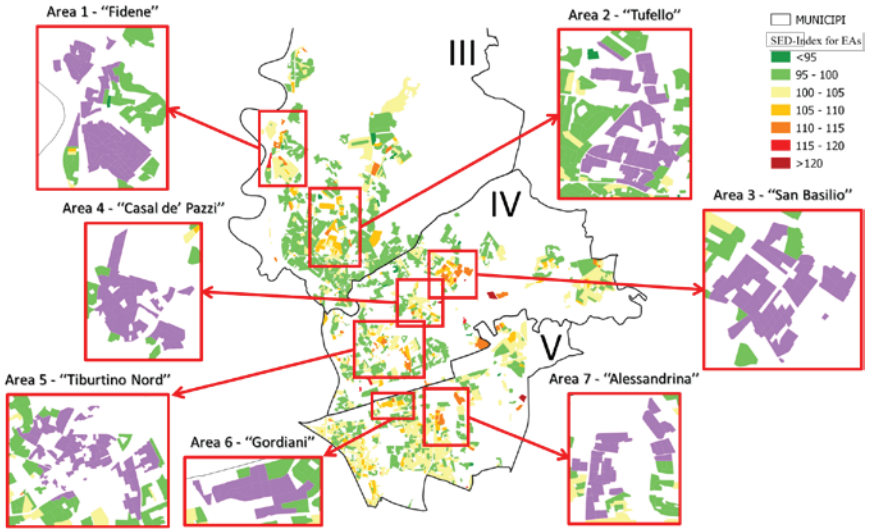
Figure 2: Map of SED-Index values on the UAs of Rome.



[4] In the Municipality of Rome, there are two hierarchical subdivisions of the territory: a less fine subdivision consisting of 15 areas called "*Municipi*" (MA), defined for administrative purposes; and a finer one that consists of 155 "*Zone urbanistiche*" (UA) defined for statistical, land planning, and management purposes.

In order to identify with greater territorial precision the areas in which deprivation is most concentrated, the spatial distribution of the SED-Index values referred to the EAs was considered. In particular, based on the evidence in Figure 2, a more detailed analysis was conducted on the MA (see footnote 4) III, IV, and V areas to the east of Rome. Figure 3 represents the relative spatial distribution of the SED-Index values on the EA map (with population and residential buildings), highlighting some areas of highest concentration of household deprivation (SED-Index>110) that can be considered critical with respect to the socio-economic phenomenon considered.

Figure 3 – Spatial classification of SED-Index values on EAs in the MA III, IV and V areas of Rome, and identification of critical areas (with the name of the UA in which the area falls).



For the seven critical areas, population, number of households, the individual indicators used in the deprivation study (Table 1) and the following indicators were calculated (Table 4): *percentage of foreigners* (Dem1); *employment rate (15+)* (Dem2); *ratio of young (0-24)/old (65+)* (Dem3); *percentage of single-member households* (Dem4); *percentage of households with 5+ members* (Dem5); *percentage of graduates (25-64)* (Dem6).

Table 4 – Population, households, individual indicators used for the study of deprivation and some demographic indicators for the 7 critical areas identified in the Rome MAs III, IV, and V.

Indicator	Rome	Area 1	Area 2	Area 3	Area 4	Area 5	Area 6	Area 7
Population	2,646,415*	14,458	22,231	9,905	9,193	16,934	7,656	9,218
Households	1,281,381*	6,584	10,902	4,578	4,446	7,917	3,762	4,363
Dep1	8.72	11.8	17.5	21.8	11.6	14.1	16.4	21.3
Dep2	11.1	9.6	10.2	11.2	10.8	8.9	9.9	10.0
Dep3	32.3	47.5	60.6	75.8	50.2	53.8	55.3	69.4
Dep4	1.1	1.8	2.0	2.6	1.6	1.8	1.9	1.8
Dep5	71.8	67.7	63.5	60.5	69.1	67.1	67.3	63.4
Dep6	20.9	30.3	33.8	46.2	28.0	33.9	30.6	37.5
Dep7	20.8	21.7	25.1	31.8	23.4	25.9	24.2	27.6
Dep8	9.5	12.8	16.2	23.5	15.4	15.0	13.4	18.3
Dep9	0.4	0.5	0.4	0.5	0.5	0.4	0.4	0.5
Dem1	12.3	15.2	6.9	4.8	18.2	6.2	12.4	7.9
Dem2	47.9	46.7	41.8	40.6	48.4	43.3	44.0	40.7
Dem3	0.9	1.0	0.7	0.9	0.9	0.8	0.8	0.8
Dem4	45.0	38.1	42.6	38.4	44.5	38.6	44.2	40.7
Dem5	3.5	3.9	3.4	4.4	3.9	4.1	3.7	3.9
Dem6	34.2	20.8	20.7	11.8	23.6	21.5	20.7	14.4

* *Figure calculated only on the 11,691 EAs involved in the study. Source: see Table 1.*

An initial reading of the results shows that the seven critical areas are characterised by lower levels of employment rate aged 35-64 (Dep5) compared to the municipal figure, educational

disadvantage (Dep6, Dep7, and Dep8) and a low presence of university graduates (Dem6). There is also a lower presence of single-member households (Dem4) and a higher presence of households with 5 or more members (Dem5) compared to the municipal value.

Turning to the analysis of the critical areas taken individually, there seems to be a sort of ranking with respect to the phenomenon of deprivation. The areas that present the most serious situations are Areas 3 and 7; in particular, Area 3, which falls in the "San Basilio" district of MA IV, presents the worst values for all deprivation indicators compared to the other areas. Area 2 suffers from high deprivation. Areas 4, 5 and 6 show similar values and can be considered to be of medium deprivation. Area 1 seems to be the one with the lowest level of deprivation compared to the others.

6. Perspectives

The project to study the socio-economic deprivation of population represents an example of the new possibilities of reading the territory through the results of the Italian Permanent Population and Housing Census and administrative archives. The results presented for the municipality of Rome were an initial experimental exercise conducted with the aim of exploring and setting up some of the possible spatial analyses that the project will allow. The definition and calculation of specific individual indicators, their synthetic measurement and spatial representation on the size of the enumeration areas made it possible to identify critical areas to focus on for further investigation.

In addition to defining the geographical limits of critical areas, it is possible to measure the different dimensions of deprivation and to study the structural characteristics of the population living in these areas by means of appropriate demographic indicators: the presence of young or elderly people, the foreign component, the level of education, employment status, and family types.

This information will be useful to better define which local policy interventions are most appropriate to support the most disadvantaged households and individuals. In this regard, the involvement of the municipalities, who know the area in which they operate and can provide the right keys to the results that will be achieved, is the most relevant aspect for the success of the project.

References

Carbonetti G., Biasciucci F., Cutillo A., Mazziotta M., Quondamstefano V., Tamburrano M.T., Tronu D. (2024a). Measuring socio-economic deprivation at sub-municipal level through the integration of census and administrative data. *Italian Journal of Economic, Demographic and Statistical Studies - RIEDS* (*forthcoming*).

Carbonetti G., Ferrara R., Gallo G., Verrascina M. (2024b). The evolution of the data production model in Istat. New perspectives for the analysis of population socio-economic phenomena. In *Proceedings of the 11th European Conference on Quality in Official Statistics (Q2024)*, Estoril, Portugal 4–7 June 2024.

De Muro, P., Mazziotta, M., Pareto, A. (2011). Composite indices of development and poverty: An application to MDGs. *Social Indicators Research*, 104, pp.1-18.

Falorsi, S. (2017). Census and social surveys integrated system. Note by the National Institute of Statistics of Italy, presented at the *UNECE/Eurostat Group of Experts on Population and Housing Censuses, Nineteenth Meeting, Geneva, Switzerland, 4–6 October 2017*.

Gallo G., Zindato D. (2018). Italy case study. In: *Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses,* UNECE New York and Geneva, 2018.

Mazziotta, M., Pareto, A. (2017). Synthesis of indicators: The composite indicators approach. In: Complexity in Society: From Indicators Construction to their Synthesis. *Social Indicators Research Series*. Filomena Maggino (Ed.), Springer, pp.159-191.

Mazziotta, M., Pareto, A. (2016). On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social Indicators Research*, 127 (3), pp.983-1003.

doi.org/10.26398/asaproc.0091

Volatility decomposition of traffic flow time series with complex seasonality using GARCH models

Maurizio Carpita[a], Rodolfo Metulini[b], Manlio Migliorati[a]
[a] Department of Economics and Management, University of Brescia, Brescia, Italy.
[b] Department of Economics, University of Bergamo, Bergamo, Italy.

1. Introduction

Accurately forecasting people's movements between specific areas is essential for urban planning, transportation optimization, and emergency management. Consequently, identifying the most effective statistical models is critical for policymakers to make well-informed decisions (Xie et al., 2020; WHO, 2022). In previous studies, we considered the problem of forecasting the traffic flow time series from September 2020 to August 2021 of "origin-destination" mobile phone hourly data, from an area with high flooding risk in the province of Brescia, Italy; in particular, for each of the 38 sub-areas of interest, the three (In, Out and Internal) flows were considered (Metulini and Carpita, 2023; Perazzini et al., 2023).

By applying a trivariate Vector AutoRegressive model with eXogenous regressors (VARX; Tsay, 2013) and Dynamic Harmonic Regression (DHR; Hyndman and Athanasopoulos, 2021) components to these time series, we found a discrete forecasting accuracy but also non-normal residuals, with heavy tails and time-varing heteroschedasticity (Carpita et al., 2024).

To explore this evidence for the 38×3 =114 residual's time series of our VARX+DHR model, the Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model (Bollerslev, 1986) has been used in this short paper. More specifically, the *multiplicative component standard* (mcs) GARCH model has been applied to the 114 residuals time series, to decompose volatility in *Daily*, *Hourly* and *Intradaily* (Engle and Sokalska, 2012).

The structure of the paper is as it follows: Section 2 describes the mcsGARCH model, Section 3 explain its estimation strategy and Section 4 presents some results and future developments.

2. The multiplicative component standard (mcs) GARCH model

Since the publication of the seminal paper published in 1982 by the 2003 Nobel Laurate Robert Engle (Engle, 1982), today the broad family of the GARCH models is well known and used in the financial literature (Bollerslev, 1986; Palm, 1996; Engle, 2001), but has rare applications in the transport field (Zhang et al., 2015; Ali et al., 2023).

Taking into account the strong hourly and daily seasonality observed in the traffic flow data, in this study we used a special version of the GARCH model, named *multiplicative component standard* (mcs), which allows to decompose the time series conditional volatility in three components (Engle and Sokalska, 2012). Considering, as in our case of study, a stationary and standardized (zero mean and unit variance) time series $\varepsilon_{t,i}$, where *t* denotes the day and *i* the hour, assuming the mcsGARCH model the *Total* conditional variance is the product of three components:

$$\varepsilon_{t,i} = \left(\sigma_t s_i h_{t,i}\right) z_{t,i} \qquad t = 1,2,..,T \quad \text{and} \quad i = 1,2,..,24 \tag{1}$$

where σ_t is a *Daily* (exogenously determined) volatility, s_i the *Hourly* volatility, $h_{t,i}$ the *Intradaily* volatility, and $z_{t,i}$ the i.i.d. standardized innovation, which conditional follows some appropriately chosen distribution. To estimate the mcsGARCH model with its conditional

variance (volatility squared) components, the two-step procedure offered by the R package rugarch by Ghalanos (2024) can be used.

In the *first step*, the Daily volatility σ_t has to be preliminarily estimated: externally from a multifactor risk model as in Engle and Sokalska (2012), or predicted from a daily GARCH model as in Andersen and Bollerslev (1997). In our study, we adopted this second approach, using the very flexible *asymmetric power* (ap) GARCH(*q,p*) model, which allows for different positive and negative effects and includes six GARCH submodels (Ding et al., 1993):

$$\sigma_t^{\delta} = \omega + \sum_{l=1}^{q} \alpha_l (|\varepsilon_{t-l}| - \gamma_l \varepsilon_{t-l})^{\delta} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta} \tag{2}$$

with $\omega > 0, \alpha_l \geq 0$, and $\beta_j \geq 0$ with the stationarity constrain $\sum_{l=1}^{q} \alpha_l + \sum_{j=1}^{p} \beta_j < 1$; the parameter $\delta > 0$ allows for the Box-Cox transformation of σ_t, and γ_l is the parameter in the leverage term (Ghalanos, 2024, sec. 2.2.5).

In the *second step*, using the Daily volatility estimated in the first step $\hat{\sigma}_t$, the Hourly variance s_i^2 is estimated by simply applying the *T*-days average of the squared daily standardized residuals:

$$\hat{s}_i^2 = \frac{1}{T} \sum_{t=1}^{T} \frac{\varepsilon_{t,i}^2}{\hat{\sigma}_t^2} \qquad i = 1,2,..,24. \tag{3}$$

Then, scaling the residuals by the estimated Daily and Hourly volatility gives $\overline{\varepsilon}_{t,i} = \varepsilon_{t,i}/(\hat{\sigma}_t \hat{s}_i)$, which are used to estimate the Intradaily component of volatility $h_{t,i}$ assuming the hourly standard (s) GARCH(*q,p*) model (Bollerslev, 1986):

$$h_{t,i}^2 = \omega + \sum_{l=1}^{q} \alpha_l \overline{\varepsilon}_{t,i-l}^2 + \sum_{j=1}^{p} \beta_j h_{t,i-j}^2 \tag{4}$$

Obtained the constrained estimates $\hat{\omega} > 0, \hat{\alpha}_l \geq 0$ and $\hat{\beta}_j \geq 0$, the persistence parameter of the hourly sGARCH model (that is a measure of how quickly past shocks decay over time assuming value in (0,1): if close to 1 volatility shocks have a long-lasting impact on future volatility) is estimated by $\hat{P} = \sum_{l=1}^{q} \hat{\alpha}_l + \sum_{j=1}^{p} \hat{\beta}_j$, and the estimated *unconditional* volatility by this model is $\hat{\sigma} = \hat{\omega}/(1 - \hat{P})$.

To estimate any GARCH(*q,p*) model, a choice must be made about the order *q* and *p* of the two lagged components on the right-hand side of its equation. As stated in Engle and Sokalska (2012) for financial data, a GARCH(1,1) model proved to be an adequate and popular choice in most of the cases and, if any higher lags are included in the conditional variance specification, the models are not of a higher order than GARCH(1,2) or GARCH(2,1). Still, efficiency considerations favor models of a lower order.

The last choice is about the conditional distribution of the standardized innovation $z_{t,i}$: for the sake of flexibility, the Normal Inverse Gaussian (NIG) is used, which allows controlling for location, scale, skewness, and kurtosis (tail heaviness) of *z*'s, and includes as a special case the Normal distribution. To reduce the computational burden, the four parameters of the NIG can be estimated from a transformation of two parameters, γ_3 and γ_4, that estimate the skewness and the shape (scale and tail) of the NIG distribution respectively (Ghalanos, 2024, sec. 2.3.5).

3. The estimation strategy of the mcsGARCH model

To estimate the three volatility components of the mcsGARCH model described in the previous section, daily and hourly time series of the 114 VARX+DHR model residuals have been used. Each time series has been split into two parts: the first 4-month period (from September to December 2020) of daily data has been used to identify the best order (*q,p*) of the

ap- GARCH model in the first step of the estimation procedure, and the last 8-month period (from January to August 2021) of hourly data has been used in the second step of the procedure to estimate the daily volatility and identify the best order (q,p) of the GARCH model for the Intradaily volatility. The hybrid solver estimation algorithm with the standardized time series was used (Ghalanos, 2024, sec. 3).

The estimation strategy adopted to apply the mcsGARCH model was as follows:

- Preliminary analysis was carried out on the 114 daily time series for the first 4-month period to assess the presence of ARCH effects ($q > 0$), using the classical Engle's Lagrange Multiplier test (Engle, 1982): the null hypothesis of variance stability (no ARCH effects) was rejected for all but 21 daily time series, therefore considered stationary in variance.
- For each of the 93 daily time series with significant ARCH effects ($q > 0$ for the Engle test), the first 4-month period was used to find the best order (q,p) of the apGARCH model, minimizing the AIC (Akaike Information Criterion) for all permutations of q in [1,2] and p in [0,1,2].
- The best apGARCH(q,p) model identified in the first 4-month period was used to forecast the daily autoregressive conditional volatility for the last 8-month period. For each of the 93 daily time series, a rolling forecast was used with 90-day length window.
- For each of the 114 hourly time series of the last 8-month period, scaled by forecast of Daily and estimate of Hourly volatility, the best order (q,p) of the sGARCH model for the Intradaily volatility was selected by the AIC for all permutations of q and p in [0,1,2].

In the end, for each of 114 time series of the VARX+DHR residuals, the total hourly volatility has been decomposed into its three components: Hourly, Daily, and Intradaily.

4. Some results and future developments

For the 93 daily time series with significant ARCH effects, the best orders (q,p) of the apARCH model (Section 3, second point) were the following: 57 (1,0), 21 (1,1), 5 (1,2), 9 (2,0) and 1 (2,2). For the 114 hourly time series, the best orders of the hourly sGARCH(q,p) model (Section 3, fourth point) were the following: 53 (1,1), 60 (1,2) and 1 (2,2).

Table 1 shows mean and standard deviation (std. dev.) of the parameter estimates for the two parts of the mcsGARCH model: results are on the left for the 93 daily apGARCH models estimated for the first 4-month period, and on the right for the 114 hourly sGARCH models estimated for the last 8-month period. Differences emerge in parameter estimates between Out and In flows versus Internal flows and between the daily and hourly GARCH models.

Table 1: Means (standard deviations) of the mcsGARCH model parameters by flow types.

	93 daily apGARCH Models (4 months)			**114 hourly sGARCH Models (8 months)**		
Parameters	*Out Flow*	*In Flow*	*Internal Flow*	*Out Flow*	*In Flow*	*Internal Flow*
Delta	1.899 (0.846)	1.986 (0.924)	2.999 (0.404)			
Gamma	-0.099 (0.341)	-0.110 (0.366)	0.484 (0.344)			
Omega	0.554 (0.334)	0.510 (0.354)	0.346 (0.229)	0.069 (0.089)	0.070 (0.080)	0.212 (0.019)
Persistence	0.564 (0.380)	0.610 (0.415)	0.841 (0.323)	0.937 (0.092)	0.937 (0.083)	0.798 (0.023)
Skew	-0.467 (0.480)	-0.480 (0.491)	-0.184 (0.190)	0.160 (0.116)	0.163 (0.128)	-0.123 (0.012)
Shape	3.165 (4.797)	3.089 (3.931)	1.707 (2.829)	4.105 (1.783)	5.124 (3.011)	1.189 (0.124)

First of all, let's consider the special parameters of the daily apGARCH model: the mean of the estimated delta is almost 2 (std. dev. almost 1) for the Out and In flows, and about 3 (std. dev. 0.4) for the Internal flows; the mean of the estimated gamma (only the first one is

considered, as only 10 time series have $q = 2$) is about – 0.1 (std. dev. about 0.3) for the Out and In flows, and about 0.5 (std. dev. about 0.3) for the Internal flows. The means for the omega (intercept) parameter estimates are much higher for the 93 daily apGARCH models (0.5 for Out and In flows and 0.3 for Internal flow, with std. dev. 0.3 and 0.2 respectively), with respect to the 114 hourly sGARCH models (0.07 for Out and In flows and 0.2 for Internal flow, with std. dev. about 0.09 and 0.02 respectively). For both Out and In flows, the mean estimated daily persistence (sum of the alpha and beta parameters) is lower (about 0.6, with std. dev. 0.4) than the hourly persistence (about 0.9, with std. dev. 0.1). In contrast, for Internal flows, persistence estimates are similar across both the two periodicities: around 0.8, but with a daily std. dev. of 0.3 and an hourly std. dev. of only 0.02. The means for the estimated skew of the NIG distribution for the GARCH residuals are negative for all daily flows (about 0.5, with analogous standard deviation) and for hourly Internal flows (lower than 0.2, and smaller standard deviation), while they are positive but very close to zero for the other two hourly flow types. The means of the estimated shape parameter of the NIG distribution are positive, ranging from 3 to 5 for Out and In flows, and from 1.2 to 1.7 for both daily and hourly Internal flows: this means that the autoregressive part of the mcsGARCH model doesn't fully capture the heavy tails observed for the VARX+DHR residual (leptokurtic) distributions.

Figure 1 is a visualized example of mcsGARCH decomposition of the estimated volatility of one of the 114 analyzed time series: under the graph of the VARX+DHR model's residual for the last 8-month period Out flow from the area of Paderno-Franciacorta, the estimated Total volatility and it's three multiplicative components (Daily, Hourly and Intradaily) are represented. In the considered period, the range of the estimated Total volatility is between 0.5 and 2, but some clusters with peaks between 4 and 6 are observed. The base level of the estimated Daily volatility is around 1, and some clusters with higher variability (up to 3) at the beginning of January, April, June, and at the end of July are observed. The estimated Hourly volatility profile, which repeats each 24 hours, shows low (below 1) variability during the night, moderate variability (about 1.5) in the diurnal part of the day, and high (peaks above 2) at 8 and 18 o'clock. The last graph for the Intradaily volatility, estimated with the hourly apGARCH model, shows smaller clusters and therefore, for this time series, can be considered as residual.

To explore relations between the traffic flows variability in the various areas under consideration, the linear correlation coefficients between the time series of the Total volatility estimates and each of its three components (93 Daily, 114 Hourly and 114 Intradaily) obtained from the mcsGARCH model were computed, and their boxplots by flow types (Out, In and Internal) are in Figure 2. Distributions for the Internal flow correlations of the Total and each estimated volatility component are very different with respect to Out and In flows ones. Considering the Total volatility estimates, the boxplots of the correlations for Out and In flows are (excluding outliers) roughly symmetric with median near 0.6 and 0.7 respectively and interquartile range about 0.1, whereas for the Internal flow is near to 1, with a very small interquartile range. Analogous correlation distributions are observed for the three estimated volatility components of the Internal flow, whereas for the Out and In flows boxplots these are lower (median between 0.3 and 0.6) for the Daily and Intradaily volatility estimates, and higher (median about 0.8 and 0.9 respectively) with negative asymmetry for the Hourly one.

This preliminary study, carried out to explore the volatility structure of the VARX+DHR model's residuals from a previous analysis, could be extended in some directions. The apGARCH model with asymmetric effects could be used to modeling the hourly volatility too, with the aim to capture with more flexibility the residual heteroskedasticity highlighted by the two parameter estimates of the NIG distribution. Moreover, volatility estimated components could be grouped using a cluster analysis approach, also to analyze the spatial distribution of traffic flows in the areas of interest. Finally, predicted autoregressive conditional volatility with a trivariate GARCH model (one equation for each flow type) for the VARX+DHR model's residuals could be used to check if the forecast of the traffic flows can be improved.

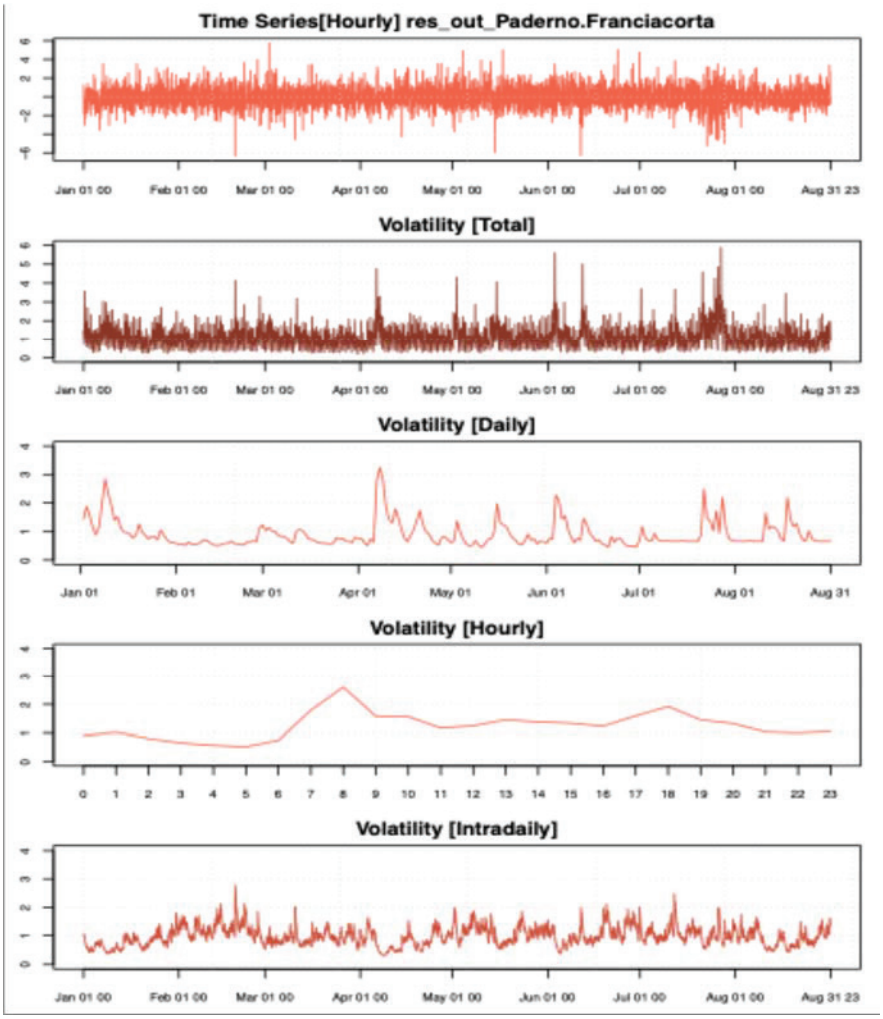


Figure 1: Example of mcsGARCH decomposition of the volatility of a time series (last 8-month period Out flow from the area of Paderno-Franciacorta). From top to bottom: The original residuals from the VARX+DHR model, the estimated Total, Daily, Hourly and Intradaily volatilities.

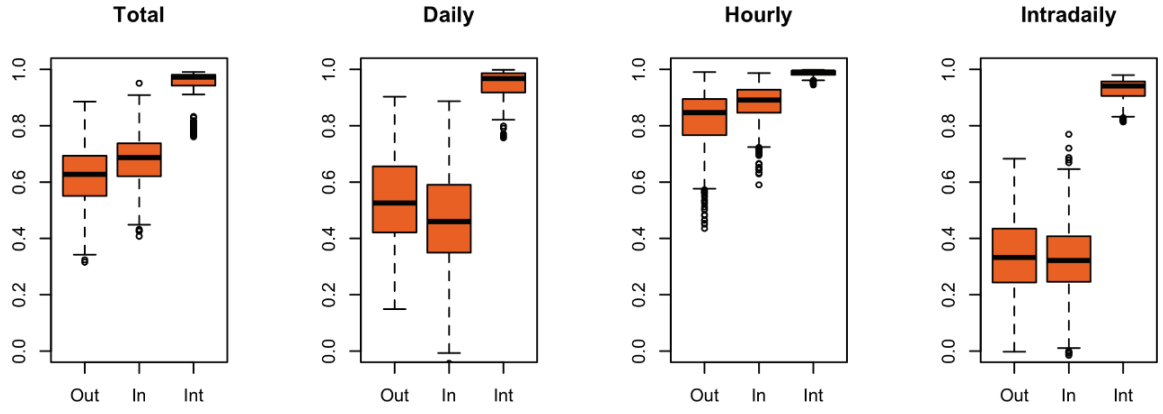


Figure 2: Boxplot of the correlations between the mscGARCH estimated Total volatility and its components (Daily, Hourly, Intradaily) by flow types (Out, In, Internal).

Acknowledgement This contribution has been developed for the project "Study of mobile phone siGNals for the evalUation of the interconnections between Mobility and the environment in Lombardia (SIGNUM)" CUP: F53D23010910001- PRIN 2022 PNRR M4C2 - financed by the European Union - Next Generation EU (DD MUR n. 1409 del 14/09/2022).
M. Carpita acknowledges the European Union (EU) and the Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), within the project "Sustainable Mobility Center (MOST)", 2022-2026, CUP D83C22000690001, Spoke N. 7, "CCAM, Connected networks and Smart Infrastructures".

References

Ali M., Yusof K.M., Wilson B., Ziegelmueller C. (2023). Traffic speed prediction using GARCH-GRU hybrid model. *IET Intelligent Transport Systems*, 17(11), pp. 2300–2312.

Andersen, T.G., Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2), pp. 115–158.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, pp. 307–327.

Carpita, M., De Luca, G., Metulini, R., Zuccolotto, P. (2024). Traffic flows time series in a flood-prone area: modeling and clustering on extreme values with a spatial constraint. *Stochastic Environmental Research and Risk Assessment*, 1–17, online first.

Ding, Z., Granger, C. W., Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1), pp. 83–106.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50 (4): pp. 987–1007.

Engle, R. F. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives*, 15(4), pp. 157–168.

Engle, R. F., Sokalska, M. E. (2012). Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *Journal of Financial Econometrics*, 10(1), pp. 54–83.

Ghalanos, A. (2024). rugarch: Introduction to the rugarch package. *R Package Version 1.4-3*.

Hyndman, R.J., and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice.* 3rd edition, OTexts: Melbourne, Australia.

Metulini, R., Carpita, M. (2023). Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. *Annals of Operations Research*, 1–26, online first.

Palm, F. C. (1996). GARCH Models of Volatility. In: Maddala, G. and Rao, C., Eds*., Handbook of Statistics,* Elsevier Science, Amsterdam, pp. 209–240.

Perazzini, S., Metulini, R., Carpita, M. (2023). Integration of flows and signals data from mobile phone network for statistical analyses of traffic in a flooding risk area. *Socio-Economic Planning Sciences*, 90, 101747, 1–17.

Tsay, R. S. (2013). *Multivariate Time Series Analysis with R and Financial Applications*. John Wiley & Sons.

World Health Organization (2022). *Urban Planning, Design and Management Approaches to Building Resilience – An Evidence Review*. WHO Regional Office for Europe, Copenhagen.

Xie, P., Li, T., Liu J., Du, S., Yang, X., Zhang, J. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 59, pp. 1–12.

Zhang, Y., Haghani, A., Zeng, X. (2015). Component GARCH models to account for sea- sonal patterns and uncertainties in travel-time prediction. *IEEE Transactions on Intelligent Transportation Systems*, 16(2), pp. 719–729, April 2015.

doi.org/10.26398/asaproc.0092

Economic well-being of households and attractiveness of territories: The case of Italian large municipalities

Sara Casacci
Department for development of methods and technologies for production and dissemination of statistical information, Italian National Institute of Statistics (Istat), Rome, Italy

1. Introduction

The integrated use of different types of data sources has become increasingly important for statistical disciplines in recent years from the point of view of socio-economic analysis. The aim of this work is to jointly investigate the economic well-being of households and the attractiveness of territories in large Italian municipalities in order to verify whether there is a relationship between the two phenomena. Economic well-being and territorial attractiveness are both multidimensional phenomena. The first one refers to the overall standard of living and quality of life of individuals and households. It encompasses various aspects such as income, wealth, employment opportunities and access to basic goods and services. Income capabilities and economic resources can be seen as the means by which an individual is able to have and maintain a certain standard of living (Istat, 2012). In advanced economies, income, employment and financial well-being are the main pillars of household economic well-being (OCSE, 2015). The attractiveness of territories is an increasingly important issue for economic development and sustainable development. In this work, the concept of territorial attractiveness is characterized by the different types of flows that a territory can attract. For example, we can have flows of investments, tourists, workers, university students (Musolino and Kotosz, 2024). There is an extensive literature on territorial attractiveness from the perspective of foreign investments (Bentivogli, 2015; Dubini, 2004), while other works focus on the attraction of tourist flows (Russo et al., 2012) or on the attraction of university students (De Angelis et al., 2017). This paper attempts to capture several aspects in relation to the following entities: university students, students at schools, workers, tourists, and enterprises. The more a territory is able to attract these entities and actively participate in them, the more it can benefit from the economic development process.

The paper is structured as follows. In Section 2, the data sources and the indicators are described and the methodological approach is presented. In Section 3, the application to the 45 large Italian municipalities (with more than 100,000 residents) is illustrated, to measure their ability to attract different types of flows in relation to the economic well-being of their inhabitants.

2. Data and method

2.1 Data

The selection of indicators was guided by the literature review. In particular, the dimensions of income and work were considered for economic well-being and the dimensions of entrepreneurship, labour market, tourism and education for attractiveness. Other indicators, such as households' wealth or housing, were not taken into account as no data was available.

The indicators used in this application come from different data sources (administrative data, statistical registers and survey data). They are:

a) the Istat microdata collection "Socio-economic conditions of households - ARCH.I.M.E.DE", which is based on the integration of several administrative sources. It collects information on socio-demographic characteristics, occupation, education and income of individuals and households in Italy.
b) The experimental Istat statistics "Day-time population for study and work", which aims to quantify and characterise an important part of the daytime population.
c) The Istat statistical register of active enterprises. It covers all enterprises that carry out economic activities in the industrial, commerce and service sectors.
d) The survey on museums and other cultural institutions. It is a census survey carried out by Istat, the Ministry of Cultural Heritage and Activities and Tourism, the Regions and Autonomous Provinces. The data collected describe over 4,500 institutes.

Table 1 contains the list of selected indicators.

Table 1. List of indicators

Theme	Code	Indicator	Description	Source
Economic well-being	X_1	Low-income households	Percentage of households with total income less than or equal to 60% of the median equivalised income	Socio-economic conditions of households - ARCH.I.M.E.DE
	X_2	Low work intensity	Percentage of households where the members of working age worked a working time equal or less than 20% of their total work-time potential during the year	
	X_3	Low-income employees	Percentage of employees with income from employment less than or equal to 60% of the median income from employment	
	X_4	Single-income households with minors	Percentage of single-income households with minors	
Attractiveness	Y_1	Coexistence index	Ratio between the daytime population[1] and the resident population in the territorial domain of interest. It quantifies the imbalance between the population using the territory and the resident population.	Experimental Statistics "Day-time population for study and work"
	Y_2	Attraction index	Ratio between the flows entering the area and the total flows generated by the territorial domain in question	
	Y_3	Entrepreneurship rate	Ratio between the number of businesses and the average resident population per thousand	Statistical register of active enterprises
	Y_4	Visitors per 100 inhabitants	Ratio between the number of visitors to museums, galleries, archaeological sites and monuments and resident population	Survey on museums and other cultural institutions

2.2 Method

A canonical correlation analysis (Hotelling, 1936) was performed[2] to measure the association between the two sets of indicators. This is a multi-dimensional data analysis method that deals with linear relationships between different groups of variables. It consists in identifying a double system of latent variables that reproduce the correlation between the two groups of observed variables, net of that present within them. In other words, starting from the original representation of the statistical units, given by the two groups of detected variables,

[1] The daytime population in a given area is composed of subpopulations of workers, students, city users, and so on. The more attractive or repulsive the area is, the more the daytime population differs from the resident population.

[2] A CANCORR procedure in SAS, which performs a canonical correlation, was run.

through this technique a synthesis is obtained. It allows representing the original variables through two new groups of artificial variables. The two groups of variables are treated symmetrically without establishing causal relationships. For these reasons, other methods that imply the assumption of causal relationships between the variables (such as multivariate regression or structural equation modelling) were not chosen.

In detail, assume that we observe two groups, respectively of k and m quantitative variables, on a collective of statistical units. The study of the correlation between the variables of one group and those of the other, would require the calculation of $k \times m$ simple correlation coefficients between the coupled variables of the two groups, of $k(k-1)/2$ and $m(m-1)/2$ simple correlation coefficients within each group, multiple correlation coefficients between the individual variables of one group and all or part of the variables of the other group.

In order to summarize these links, the canonical correlation analysis uses a double system of latent variables, which reproduce the correlation between the two groups of original variables, net of that existing within them. In practice, two new groups of variables are constructed that are uncorrelated within themselves and maximally correlated with each other. The latent variables thus determined are called canonical variables. In each of the two groups, the canonical variables are constructed as a linear combination of the manifest variables, with weights expressed by the link of each manifest variable with the latent phenomenon that the canonical variable represents (Bouroche and Saporta, 1983). These are obtained simultaneously for one and the other group, in such a way as to maximize the variability within each group and the correlation between the two groups. The first pair of canonical variables is the one that shows the greatest correlation; the correlation between successive pairs decreases as they are determined.

3. Results

In this section, the results of the proposed approach are illustrated. They refer to the 45 large Italian municipalities in 2019. The main summary statistics for each indicator – minimum, mean, maximum and standard deviation – are presented in Table 2.

Table 2. Descriptive statistics for X_1-X_4 and Y_1-Y_4

Statistic	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	Y_4
Min	14.6	16.30	3.10	21.5	95.03	20.45	52.66	2.66
Mean	24.89	26.12	6.86	30.11	123.20	39.76	87.34	382.34
Max	50.7	44.3	15.4	39.9	154.94	57.57	135.35	3184.81
Std. Dev.	9.35	8.04	3.33	4.50	14.65	9.34	19.94	567.23

The analysis of the canonical correlation shows that the correlation between the first pair of canonical variables is equal to 0.744. The hypothesis test shows that the probability that the canonical correlation is zero is 0.0010, so the first canonical correlation is statistically significant. The other canonical correlations are lower and are not statistically significant; therefore, they are not considered (Table 3). A sensitivity analysis was also performed to test the robustness of the results by varying the selection of indicators. In particular, each indicator was iteratively excluded from the analysis and the canonical correlation procedure was performed. For each selection of indicators, the results showed a high and statistically significant correlation between the first pair of canonical variables. Only the removal of the entrepreneurship rate leads to a reduction in the canonical correlation. The inclusion of alternative indicators was also tested. The analysis results were still robust, but the contribution of alternative indicators to the increase in canonical correlation was low. Based on the results obtained, the first pair of canonical variables is selected. The canonical variable

of the first group represents the factor "economic well-being of households", while the canonical variable of the second group of indicators summarizes the dimension "attractiveness".

Table 3. Canonical correlation analysis and test of hypothesis result

Pair of canonical variables	Canonical correlation	Adjusted canonical correlation	Approximate standard error	Squared canonical correlation
1	0.744077	0.701534	0.068873	0.553651
2	0.433019	0.351156	0.125371	0.187506
3	0.145765	.	0.151025	0.021247
4	0.10516	.	0.152597	0.011059

Test of H0: The canonical correlations in the current row and all that follow are zero

Pair of canonical variables	Likelihood Ratio	Approximate F value	Num DF	Den DF	Pr > F
1	0.3510254	2.75	16	107.56	0.001
2	0.78643691	1.01	9	87765	0.437
3	0.96792897	0.3	4	74	0.8744
4	0.98894145	0.42	1	38	0.5184

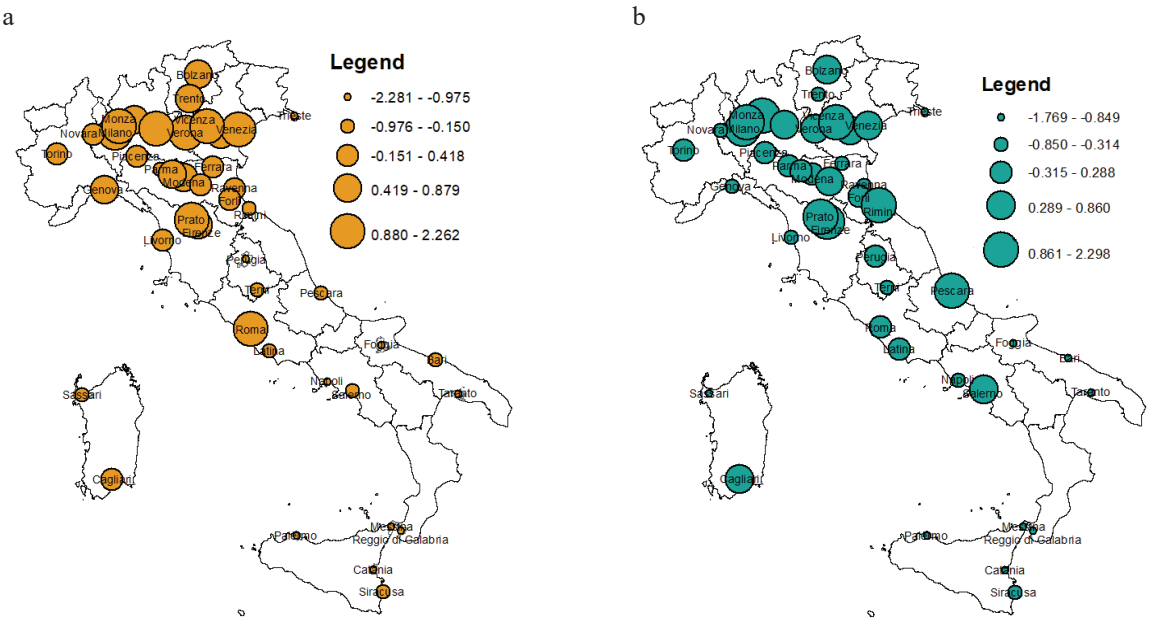
Table 4 report the correlations between the indicators of the first group and the corresponding first canonical variable and the correlation between the indicators of the second group and the corresponding first canonical variable. The first canonical variable of the attractiveness group is strongly correlated with the entrepreneurship rate, and it is quite correlated with the attraction index and coexistence. In contrast, it is less correlated with the number of visitors per 100 inhabitants. As regards economic well-being, the contribution of 'low-income households' and 'low work intensity' to the first canonical variable is high.

Table 4. Correlations between indicators of the two groups and respective canonical variable

Economic well-being indicators	Correlation	Attractiveness indicators	Correlation
Low-income households	-0.7012	Attraction index	0.5248
Low work intensity	-0.7688	Coexistence index	0.5046
Low-income employees	-0.6068	Entrepreneurship rate	0.9440
Single-income households with minors	-0.4738	Visitors per 100 inhabitants	0.3478

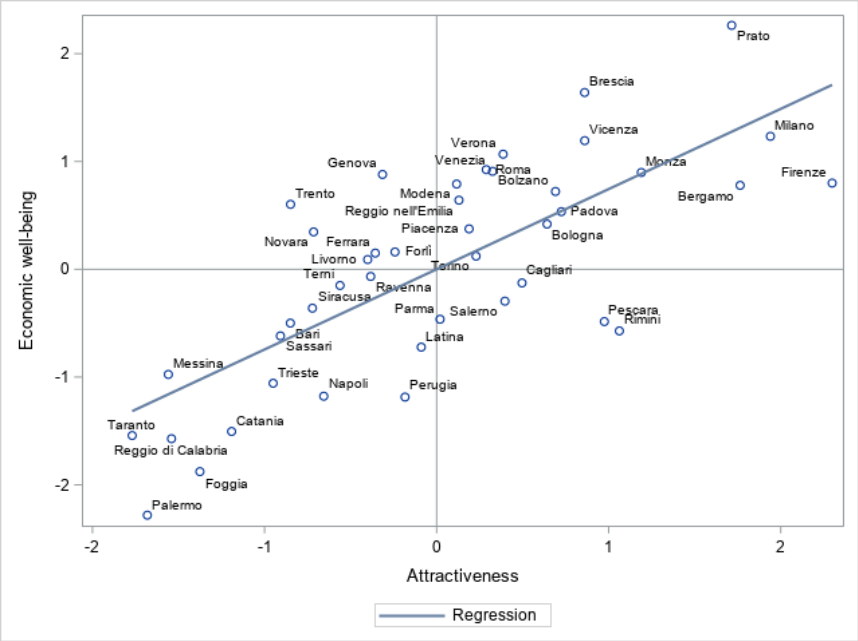
The values assumed by the two canonical variables thus identified are associated with each of the statistical units (the large municipalities) in order to analyse these two new "synthetic" variables. The cartograms of the territorial distribution of the two canonical variables (Figure 1) highlight the North-South dualism in the two dimensions. In particular, the municipalities of Lombardia, Veneto, Trentino Alto Adige and Emilia Romagna present very high values on the first canonical variable underlying economic well-being. In the Centre, Roma, Firenze and Prato show high values, while in the South, economic well-being is generally very low. As far as attractiveness is concerned, very low values are observed in Puglia, Calabria and Sicilia. These results are consistent with the literature, which shows that the Southern Italian regions have lost productive factors (human capital, entrepreneurs, etc.) and suffer from phenomena such as brain drain, which have affected their economic development potential (Vecchione, 2017).

Figure 1. Large Italian municipalities by economic well-being (a) and attractiveness (b) - Canonical variables (quintiles)



The relationship between the dimensions is illustrated in Figure 2, which shows the scatter diagram for the two canonical variables. The upper right quadrant of Figure 2 contains the municipalities with the highest levels of economic well-being and attractiveness. Prato, Milano, Brescia, Vicenza, Monza, Bergamo and Firenze stand out here. Prato is based on the textile industry, with one of the largest districts in Europe. Brescia, Bergamo, Milano and Monza are among the main productive centers in Italy. Milano and Firenze are very attractive for tourism and are home to universities attracting many off-site students (Vivio, 2016).

Figure 2. Large Italian municipalities by economic well-being and attractiveness



The cities of the Islands and the South have low values in both dimensions and are mainly located in the lower left quadrant, with Palermo, Reggio Calabria, Taranto and Foggia having the lowest values. The presence of some notable peculiarities is highlighted. This is the case

of Pescara and Rimini, which have quite high values in the attractiveness dimension and below average values in the economic well-being dimension. In particular, Pescara has a high coexistence index; Rimini is a tourism-oriented economy but ranks last place in Emilia-Romagna in terms of family income. On the contrary, Genova and Trento have a fairly high level of economic well-being but are not very attractive.

4. Conclusion

The joint analysis of the economic well-being of households and the attractiveness of territories in large Italian municipalities proposed in this paper underlines the great potential of integrating indicators from different types of data (surveys, integrated administrative data, and statistical register). In particular, two set of indicators are studied using a multi-dimensional data analysis method. The results show a rather strong relationship between them. They also confirm a North-South dualism and have some remarkable peculiarities. The perspectives of analysis point to the possibility of providing useful knowledge to support decision-makers and the implementation of policies for specific areas. Looking at economic well-being and attractiveness at the local level provides useful insights for policy makers and facilitates the assessment of the effectiveness of strategies to reduce territorial disparities. To promote attractiveness and improve economic well-being in lagging regions, policy frameworks need to address disparities, improve access to public services and infrastructure, and increase productivity and competitiveness.

References

Bentivogli, C., Cherubini, L., Iuzzolino, G. (2015). L'attrattività dei territori: i fattori di contesto. *Politica Economica*, **1**, pp. 3-26.

Bouroche, J.-M., Saporta, G. (1983). *L'analyse des données*. C.L.U. Editrice, Napoli.

De Angelis, I., Mariani, V., Torrini, R. (2017). New evidence on interregional mobility of students in tertiary education: The case of Italy. *Politica Economica*, **1**, pp. 173–96.

Dubini, P. (ed.) (2004). *L'attrattività del sistema Paese. Profili territoriali e idealtipi per le imprese*. Il Sole 24 Ore, Milano.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28** (3-4), pp. 321-377.

Istat (2012). *Descrizione dei domini e degli indicatori del Bes selezionati dalla Commissione scientifica e varati il 22 giugno 2012*. https://www.istat.it/it/files/2018/04/12-domini-commissione-scientifica.pdf.

Musolino, D. Kotosz, B. (2024). A new territorial attractiveness index at the international scale: Design, application and patterns in Italy. *The Annals of Regional Science*, **72**, pp. 1159-1187.

OCSE (2015). *Households' Economic Well-Being: The OECD Dashboard Methodological Note*. https://www.oecd.org/sdd/na/OECD-Household-dashboard-methodological-note.pdf.

Russo, A., Smith, I., Atkinson, R., Servillo, L., Madsen, B., van der Borg, J. (2012). *ATTREG. TheAattractiveness of European Regions and Cities for Residents and Visitors. Final Report*. Tarragona, ESPON & University Rovira i Virgili.

Vecchione, G. (2017). Migrazioni intellettuali ed effetti economici sul Mezzogiorno d'Italia. *Rivista Economica del Mezzogiorno*, **3**, pp. 643-662.

Vivio, R. (a cura di) (2016). *Studenti e bacini universitari.* Istat, Roma.

doi.org/10.26398/asaproc.0093

An application of the partial credit model to measure the resilience of Italian young adults in the aftermath of the COVID-19 pandemic

Clelia Cascella[a], Luigi Bollani[b], Luigi Fabbris[c]
[a] Department of Test Development, INVALSI, Rome, Italy.
[b] Department of Economics, Social Studies, Applied Mathematics and Statistics, University of Turin, Turin, Italy.
[c] Tolomeo Studi e Ricerche, Padua, Italy

1. Introduction

The concept of resilience has been widely discussed in the literature, in particular with reference to the positive psychological capital of persons (Seligman and Csikszentmihalyi, 2000; Luthans et al., 2007; Fletcher and Sarkar, 2013). Although there is still no consensus on its definition, we can define resilience as a mental construct that plays an important role in helping individuals to cope with life's challenges and to bounce back from adversity.

Resilience has therefore always played a central role in personal growth and fulfilment. Such a role has become even more important during and after the Coronavirus spread. The pandemic has had a significant impact on people's mental health. In addition, the strict lockdown measures and the shift to remote working and/or learning exacerbated feelings of isolation and anxiety, leading to increased levels of depression and other psychopathologies, in particular among Italian youngsters (Serafini et al., 2020; Bollani et al., 2023). Measuring resilience should therefore be considered a priority for the social research agenda.

In this paper, we propose and apply a resilience scale on a sample of Italian adults and highlight the specificity of young adults (aged 18-34). Our scale is composed of (only) 8 items and aims to overcome the unaffordable burden posed on respondents by self-administering long scales. Indeed, ours is a reduction of the 10-item scale by Campbell-Sills and Stein (2007), which, on turn, is a reduced form of the 25-item CD-RISC test by Connor and Davidson (2003).

The rest of the paper is organised as follows: Section 2 presents an overview of data at hand and of the model for data analysis. Section 3 focuses on results of the application of a resilience test to adults in Italy. In particular, we show how resilience is influenced by demographic factors and compare resilience levels among young Italians with those observed among older individuals. The implications of our research for both the policy and research agendas are discussed in the final section (Section 4).

2. Data and analytic model

2.1 Data

The sample consists of 1,192 Italian adults who took a self-administered questionnaire following the global pandemic caused by the SARS-CoV-2 virus (Bollani et al., 2002). The sample is gender balanced. Young adults ([18-34)], i.e. our target age group, represent almost half of the sample (33% [18-24]; 14% in [25-34]; 16% in [35-49], 28% in [50-64], and the remaining 9% were older than 64 years. The following list presents the items used in the present paper:

1. I can deal with any problem that comes my way in life.
2. I try to see the positive side of problems.
3. By coping with stress, I can become stronger.
4. I tend to recover after illness, injury or adversity.
5. I can achieve results even when I encounter obstacles.

6. I can stay focused and think clearly even when I am under pressure.
7. I think of myself as a strong person who can cope with life's challenges.
8. I can handle unpleasant or painful feelings.

Each item presents five answer options, ranging from 1 (indicating a strong disagreement) to 5 (indicating a strong agreement). The set of administered items encompasses a range of situations, varying in difficulty. The easiest situations are those that are more likely to be endorsed, while the most difficult ones are those that are less likely to be agreed upon.

2.2 Analytic model

To validate the proposed item set, we used Winsteps (Linacre, 2023) to estimate both a rating scale model (RSM; Andersen, 1977) and a partial credit model (PCM; Masters, 1982), both of which are extensions of the Rasch model (1960/1980) for polytomous items, with either the same or a different rating scale structure. Comparing the results of these two models was part of our validation strategy. In fact, although our selected items have the same scoring structure (suitable for the RSM), we did not take the adequacy of the RSM for granted but compared the two sets of item difficulties and person abilities estimated via the RSM and the PCM. For rating scales with the same number of categories, the total statistical information for an item is the same for the partial-credit and rating-scale models. Nonetheless, RS and PC model specification is not the same and controlling for the consistency of results from one of the other models is recommended to control for construct validity (Linacre, 2000): if results change, which analysis is more meaningful? Which is the item hierarchy that better represent the construct that a certain set of items is intended to measure? In the present study, the results of the two analyses overlapped completely, further confirming the item difficulty hierarchy and thus the construct validity of our proposed scale.

All Rasch-family models are particularly suited to the task of scale validation (Baghaei, 2008). Items that do not fit the Rasch model may be indicative of multidimensionality and could potentially be modified, discarded, or suggest that our construct theory may require further refinement. It may be reasonably assumed that items which fit the Rasch model are measuring the single dimension intended by the construct theory. The Rasch model allows the construction of a hypothetical unidimensional line, along which items and individuals can be positioned in accordance with their respective difficulty and ability measures. Items situated in close proximity to the hypothetical line contribute to the measurement of the single dimension delineated in the construct theory, while the items which fall far from the hypothetical line are measuring another dimension that is irrelevant to the main Rasch dimension. Misfitting items can thus be considered as indicative of construct-irrelevant variance as well as gaps along the unidimensional continuum are indicative of construct under-representation. The methodological strategy deployed to validate our scale was described with more details in Cascella et al. (2024).

Along such a hypothetical unidimensional line, respondents are scaled depending on the amount of resilience that is necessary to agree with each single item: respondents showing the highest levels of resilience are positioned at the top of the line as well as the items that are perceived as more difficult to be endorsed, and vice versa. The respondents, who were scaled along with the same unidimensional line, can be directly compared because, whether the model's assumptions are met, the model's estimates are invariant measures of the construct (i.e. resilience) that the proposed scale is designed to measure (Engelhard, 2013).

Scale validation carried out within the framework of the Rasch analysis is an iterative process within which the researcher explores data-model fit, thus also exploring the contribution of each item to the measurement of the construct that the developing scale purports to measure. In doing so, the validation process identifies the set of items that best represents the population of possible items, and thus the construct itself: scale validation is therefore not just a statistical exercise, but also (and foremost) a support to the definition of the construct of interest.

3. Results

Data model fit. Here, we report just the data-model fit statistics (Table 1). Ideal value of both infit and outfit MNSQ is 1.00, with a tolerate deviance of ±0.20 (Enghelhart, 2013). The analysis shows that our reduced 8-items scale does fit the Rasch model, thus making Rasch estimates invariant across sub-groups of respondents. When data fit the model's assumptions, then estimated parameters are invariant measures. Measurement invariance is a key statistical feature of the model as it allows for robust comparisons between sub-groups of respondents.

Table 1. Items statistics, ordered by Infit MNSQ size (n = 1,192)

Item label	Item Rasch parameter	Model Standard Error	Infit MNSQ	Outfit MNSQ	Point-biserial correlation
ITEM_G	0.37	0.04	1.18	1.19	0.68
ITEM_C	-0.09	0.04	1.16	1.17	0.70
ITEM_D	0.44	0.04	1.14	1.14	0.69
ITEM_J	0.21	0.04	1.06	1.08	0.68
ITEM_I	-0.03	0.04	0.96	0.99	0.73
ITEM_E	-0.56	0.04	0.91	0.88	0.68
ITEM_B	0.08	0.04	0.79	0.80	0.75
ITEM_F	-0.42	0.04	0.73	0.73	0.73
Mean	*0.00*	*0.04*	*0.99*	*1.00*	
SD	*0.33*	*0.00*	*0.16*	*0.16*	

Note: PERSON: Separation: 2.29; Reliability: 0.84 | ITEM: Separation: 7.44; Reliability: 0.98
Source: Our adaptation from Cascella et al. (2024)

Differential Item Functioning (DIF) analysis. In order to better explore measurement invariance, Table 2 reports the results of a DIF analysis (Osterlind and Everson, 2011) by age, as this is the focus of our analysis[1]. For each item, in addition to statistical significance, we also reported on DIF size that expresses the magnitude of differential functioning. In interpreting results, we align with the psychometrical literature (Zwick, 2012) according to which, even if statistically significant, DIF size can be considered (1) negligible, when lower than 0.43, (2) moderate, when it runs between 0.44 and 1.00, or (3) severe, when bigger than 1.00. Negligible values are considered not disruptive for measurement and thus do not raise concerns in terms of scale validity. Just few values reported in Table 2 are statistically significant but none of them are above |0.43|. All the differences by age are thus robust and can be interpreted from a substantive point of view.

Table 2. DIF statistical significance and size (n = 1,192)

	age groups				
Item	18-24	25-34	35-49	50-64	64+
I can deal with any problem that comes my way in life	-0.19*	-0.27*	0.14	0.23*	0.20
I try to see the positive side of problems	0.24*	-0.09	0.00	-0.14	-0.43*
By coping with stress I can become stronger	0.07	-0.23*	-0.02	0.00	0.13
I tend to recover after illness, injury or adversity	-0.11	0.22	0.04	0.02	-0.08
I can achieve results even when I encounter obstacles	-0.22*	0.14	0.06	0.13	0.16
I can stay focused and think clearly even when I am under pressure	0.17*	0.27*	-0.10	-0.19*	-0.38*
I think of myself as a strong person who can cope with life's challenges	0.06	-0.05	-0.08	-0.05	0.10
I can handle unpleasant or painful feelings	-0.06	0.00	-0.03	0.00	0.28

Note. The presence of an asterisk (*) adjacent to a reported value in the table denotes a statistically significant DIF.

Having verified that the DIF analysis does not raise concerns about measurement invariance by age, the graphs presented in Figure 1 shows the distribution of the estimated person parameters (resilience) in each age group.

By comparison, it is easy to see that - although person parameters are on the same scale - the range within which resilience is estimated varies greatly by age. Furthermore, if we look at the distribution of respondents in each age group, we can see that resilience increases with age: the older the respondent, the higher the estimated resilience.

[1] A comprehensive differential item functioning (DIF) analysis, which accounts for age and other variables identified as critical in the existing literature, has been conducted and documented in Cascella et al. (2024).

For example, in the 18-24 age group, 100 out of 390 respondents (i.e. roughly 25%) is scaled below 0.00 (where 0.00 indicates the mean of the resilience distribution on the Rasch scale). We observe pretty much the same percentage in the next age group of young adults [24-34]. However, this value drops to 0.04% in the [35-49] age group and to around 12% in the [50-64] and [64 and more] age groups, thus indicating that the proportion of respondents scaled above 0.00 on the estimated Rasch scale (i.e. the proportion of respondents that shows moderate to high levels of resilience), increases by age.

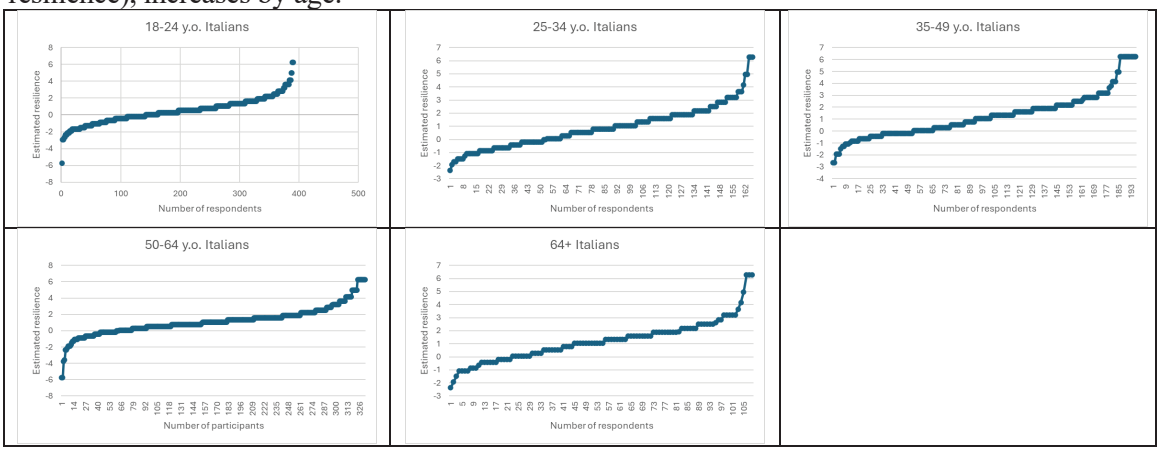


Figure 1. The distribution of estimated person parameters (resilience) by age groups

In order to better explore such a result, we use another notable feature of the Rasch model. In fact, the Rasch model allows ordinal data to be transformed and scaled into continuous measures that can be used for further statistical analysis. Once the measurement invariance of the estimates has been checked, the estimated parameters can thus be compared by the combination of different variables. We used this feature of the model to better understand the observed difference between younger and older respondents. In particular, we looked at the variability of resilience at the intersection of age and gender. This is only one of the possible combinations suggested by the literature. Results presented in Table 3 should therefore be taken as an example of what the model highlighted. In reading results, it is worth noting that, as a rule of thumb, a 0.5 difference in both item and person parameters is considered a meaningful difference. The difference between men and women in the first age group ([18-24 y.o.]) is 0.763, in favour of boys. This difference decreases from the first to the next age intervals, i.e. the gender gap decreases with increasing age, until it reverses for those aged 64 and over. This trend is in line with the literature (e.g. Sanzò et al., 2023).

Table 3. Estimated resilience at the intersection of gender and age (n = 1,192)

	18-24 y.o.			25-34 y.o.			35-49 y.o.			50-64 y.o.			64+			
	Overal l	M	F	Overal l	M	F	Overal l	M	F	Overal l	M	F	Overal l	M	F	Tota l
Estimated resilience	0.44	0.94	0.18	0.89	1.26	0.67	1.25	1.73	0.98	1.17	1.31	1.03	1.21	1.10	1.36	0.91
N	390	132	258	166	61	105	196	70	126	332	159	173	108	61	47	1192

4. Discussion and conclusion

By fostering resilience, individuals can increase their ability to overcome adversity and achieve personal growth. The results presented in this paper, based on the application of a new resilience

scale validated using Rasch modelling provided new insights into the characteristics of the most and least resilient individuals in a sample of Italian respondents who completed our questionnaire in the aftermath of the Covid-19 pandemic.

The data analysis presented in this paper showed that young adults are most at risk of achieving low or even very low levels of resilience. The proportion of respondents at risk becomes even larger when considering gender in combination with age: young Italian women emerge as the most vulnerable category. Such a result is in line with previous studies showing that young women tend to achieve lower levels of resilience than (young) men.

Our research contributes to knowledge in at least two main ways. First, our proposed scale can be seen as a tool to describe the actual levels of resilience in Italy after Covid-19, thus supporting reflections on the reasons why Italian young adults are less resilient than older ones. Second, the proposed methodology, together with the substantive results presented in the paper, can be seen as a tool to inform and channel policies and practices: in the paper we have shown that Rasch modelling can guarantee measurement invariance, thus allowing the identification of personal factors associated with different levels of resilience.

Results presented in the current paper are the preliminary part of a larger study currently under review (Cascella, Bollani, Fabris, under review). Further analysis, such a sensitivity analysis in addition to the DIF analysis we have already carried out, may further help the reader to understand how each item function across different sub-groups of respondents thus shedding new further light in the measurement and in the understanding of resilience in our contemporary society.

What we have proposed in the paper is a methodological framework within which further research can be carried out, thus supporting the advancement of knowledge in this area, which can be crucial in order to incorporate, into (health, social and educational) policies and practices, measures that support people in strengthening their resilience.

References

Andersen, E. B. (1997). The rating scale model. In *Handbook of Modern Item Response Theory* (pp. 67-84). New York, NY: Springer New York.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction*, 22 (1), pp. 1145-1146.

Bollani, L., Di Zio, S., Fabbris, L. (2023). After-COVID remote working and towards-the-future disposition in Italy. In L. Fabbris, S. Mignani, G. Tassinari (eds). *Statistics, Technology and Data Science for Economic and Social Development. Book of Short Papers of the ASA Bologna Conference, Statistica Applicata – Italian Journal of Applied Statistics*, 35(3), pp. 57-62.

Campbell-Sills, L., Stein, M.B. (2007). Psychometric analysis and refinement of the Connor-Davidson Resilience Scale (CD-RISC): Validation of a 10-item measure of resilience. *Journal of Traumatic Stress*, 20(6), pp. 1019-1028.

Cascella, C., Bollani, L., Fabbris, L. (2024). Measuring persons' resilience: Properties of an Italian inventory scale. Submitted to *Social Indicators Research*.

Connor, K.M., Davidson, R.T. (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC), *Depression and Anxiety*, 18, pp. 76–82.

Engelhard Jr, G. (2013). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*. Routledge.

Fletcher, D., Sarkar, M. (2013). Psychological resilience: A review and critique of definitions, concepts, and theory. *European Psychologist*, 18(1), pp. 12-23.

Linacre, J.M. (2023). *Winsteps® (Version 5.6.0)* [Computer Software]. Portland, Oregon:

Winsteps.com. Available from https://www.winsteps.com/.

Luthans, F., Avolio, B.J., Avey, J.B., Norman, S.M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology,* 60(3), pp. 541–572.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, pp.149-174.

Osterlind, S.J., Everson, H.T. (2011). *Differential Item Functioning*, 2nd edition online. Sage. DOI: https://doi.org/10.4135/9781412993913

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Achievement Tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Sanzò, S., Borgonovo, F., Bernardelli, G., Delle Fave, A. (2023). Mental health and daily experience of Italian college students during the COVID-19 pandemic. In *Emerging Adulthood in the COVID-19 Pandemic and Other Crises: Individual and Relational Resources* (pp. 55-75). Springer.

Seligman, M.E.P., Csikszentmihalyi, M. (2000). Positive psychology. *American Psychologist*, 55(1), pp. 5–14.

Serafini, G., Parmigiani, B., Amerio, A., Aguglia, A., Sher, L., Amore, M. (2020). The psychological impact of COVID-19 on mental health in the general population. *QJM: An International Journal of Medicine*, 113(8), pp. 529-535.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 12-08.

doi.org/10.26398/asaproc.0094

Data science approaches to the study of multimorbidity patterns at older ages

Niccolò Caselli , Linh Hoang Khanh Dang , Rosella Rettaroli , Rossella Miglio .
Department of Statistical Sciences, University of Bologna, Bologna, Italy.

1. Introduction

One key factor to construct appropriate and sustainable healthcare plans for aging populations is the possibility of identifying precise multimorbidity patterns and seizing their progression in time, so that groups of individuals with different degrees of vulnerability can be recognized and provided with assistance in a timely manner. The primary objective of this work is to characterize the complex network of multimorbidities among older adults in Europe identifying common multimorbidity clusters and their associations with demographic variables such as age and gender. The analysis utilizes data from the Survey on Health, Aging and Retirement in Europe (SHARE), wave 7 and wave 9, combining empirical data with innovative methodological approaches, based on the application of Mixed Graphical Models (MGM). MGM networking methods excel in identifying both direct and indirect relationships between multimorbidities considering the entire network of conditions simultaneously. This can reveal dependencies and interactions that may not be apparent through conventional statistical analyses.

2. Data and methods

Data from SHARE (Bergman et al, 2019, 2024, Börsch-Supan 2013), wave 7 and wave 9 are utilized to analyse multimorbidity among individuals aged 50 and older. Our samples include 72,827 and 65,772 subjects aged 50 and older, across European countries.

The study focuses on collected physical health variables, categorized as present or absent, and explores their prevalence across gender and age groups. In both waves the final list of considered chronic diseases comprises: Heart Attack, Stroke, Chronic Lung Disease, Cancer, Stomach Ulcer, Parkinson's Disease, Hip/Femoral Fracture, Alzheimer's Disease, Rheumatoid Arthritis, Osteoarthritis, Chronic Kidney Disease, Obesity (through BMI), Hearing Impairments, Hyperlipidaemia, Hypertension, Diabetes, Vision Impairments and Depression.

The dataset was segmented by age and sex into four groups: males aged 50-64, females aged 50-64, males aged 65+, and females aged 65+.

Mixed Graphical Models (MGMs) were adopted to uncover relationships among comorbidities, offering new insights into health patterns in the aging population across Europe. The comprehensive nature of SHARE data and its longitudinal design make it an ideal resource for studying these complex interactions.

Weighted, undirected pairwise chronic diseases networks are constructed with MGMs (Yang et al., 2014). The methods identify nodes; in our case each node represents one chronic disease, coded as a binary variable. The association analysis is limited to diseases dyads (i.e., mixed graphical models of interaction at order 2).

The undirected network of chronic diseases presented in this study is the network of pairwise partial correlations between diseases estimated using pairwise mixed graphical model in a generalized linear model framework. The original idea for this approach comes from the Ising model of magnetic behaviour of metals (Ising, 1925), wherein the objects (or variables) can interact with each other but only with their direct neighbours (i.e., the set of nodes that are

directly connected to the node in consideration). Following this configuration, the estimation of the chronic diseases network consists of two steps, including the identification of the best neighbour for each variable, and the determination of the presence or not of an edge between two nodes. In the first step, we estimate the network structure using nodewise regression by considering each variable of chronic disease as the response variable and all other variables as predictors. Rather than relying on an arbitrary cut-off or multiple significance tests, a combination of LASSO techniques and EBIC is applied to each regression to optimize the neighbourhood selection for each node and obtain a network more apt for interpretation purpose. In the second step, we obtain two regression coefficients for each edge and follow a pre-determined rule to decide whether an edge is present or absent. If an edge is identified to be present, we define its weight as the mean of the two regression coefficients linked to the edge. This weight is then represented as the thickness of the edge in the resulted network and describes the correlation strength between two corresponding nodes. Mixed graphical model estimations are done using package mgm (version 1.2-14) in R (Haslbeck and Waldorp, 2020). Network structure visualization and network centrality measures are made using package qgraph (version 1.9.5) in R (Epskamp et al., 2012).

Community detection algorithm to group chronic diseases into clusters within the network structure estimated by graphical models were considered adopting the Question-Alignment approach (Smith et al., 2020). This approach permits to identify the most suitable community detection algorithm for our research question and data type. The set of most common algorithms to choose from is available in R package igraph (Csardi and Nepusz, 2006).

3. Results

For the sake of space, the results are shown only for one age class, separately for male and female and wave 7. The outcomes obtained in wave 9 are consistent with those of wave 7.

Figure 1 shows the results of the MGM application considering all diseases in males aged 50-64. The three different choosen clustering methods, Fast-greedy, Louvain and Optimal algorithm, produced the following clusters represented with different colours. RED cluster encompasses cardiovascular and metabolic diseases, YELLOW cluster, which is the most heterogeneous, includes some chronic diseases. The GREEN cluster is formed by Alzheimer's disease and vision impairments. The GRAY cluster emphasizes the connection between hearing impairments and depression. The PURPLE, LIGHT BLUE, and DARK BLUE clusters are characterized by single diseases: Hip/Femoral Fracture, Cancer, and Parkinson's disease respectively.

In this network, the strongest relationships are observed between Stroke and Alzheimer, Vision Impairments and Alzheimer, Stroke and Hypertension, Heart Attack and Hyperlipidaemia, Rheumatoid Arthritis and Osteoarthritis, Stomach Ulcer and Chronic Kidney Disease (CKD), and Stomach Ulcer and Chronic Lung Disease. The MGM model demonstrated satisfactory predictive performance, with classification accuracy ranging from 0.71 to 0.99.

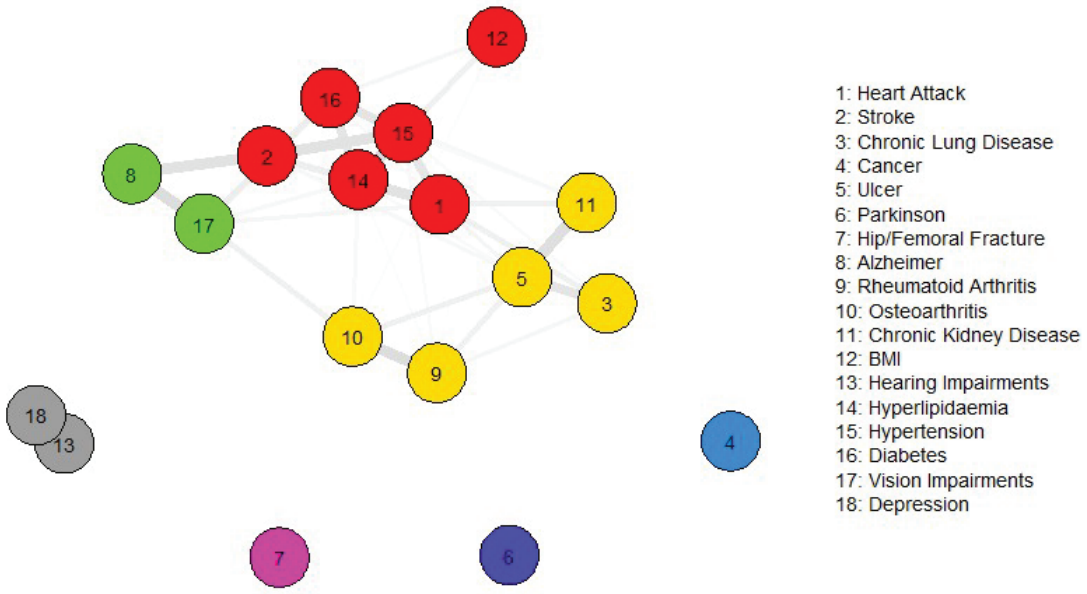


Figure 1. MGM for males aged 50-64: All diseases, SHARE Wave 7.

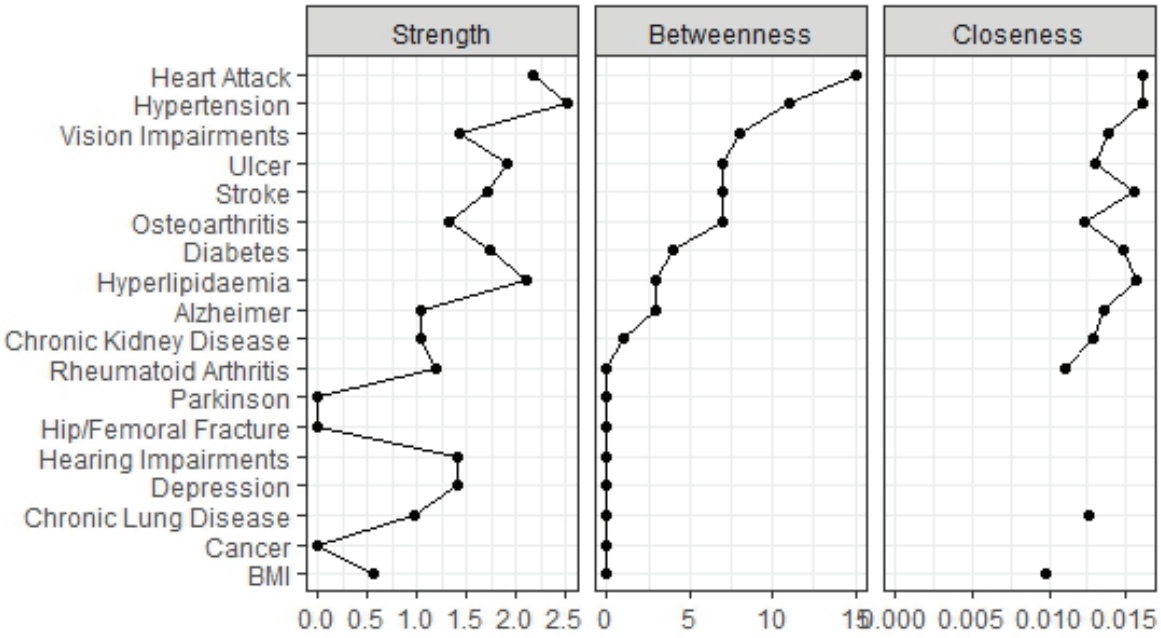


Figure 2. Centrality plot for males aged 50-64: All diseases, SHARE Wave 7.

Centrality measures such as strength, betweenness, and closeness are used to identify key nodes in a network. Figure 2 shows the centrality plot, the highest strength centrality nodes are related to cardiovascular diseases: Heart Attack, Hypertension, and Hyperlipidaemia. Heart Attack and Hypertension also show the highest betweenness centrality, acting as network bridges. The highest closeness centrality is seen in Heart Attack, Hypertension, Stroke, and Hyperlipidaemia.

Nodes in the PURPLE, LIGHT BLUE, DARK BLUE, and GRAY clusters exhibit no closeness centrality, indicating weak or isolated positions in the network.

Overall, results obtained by considering all diseases, reported that the most recurrent clusters were those associated with cardiometabolic diseases.

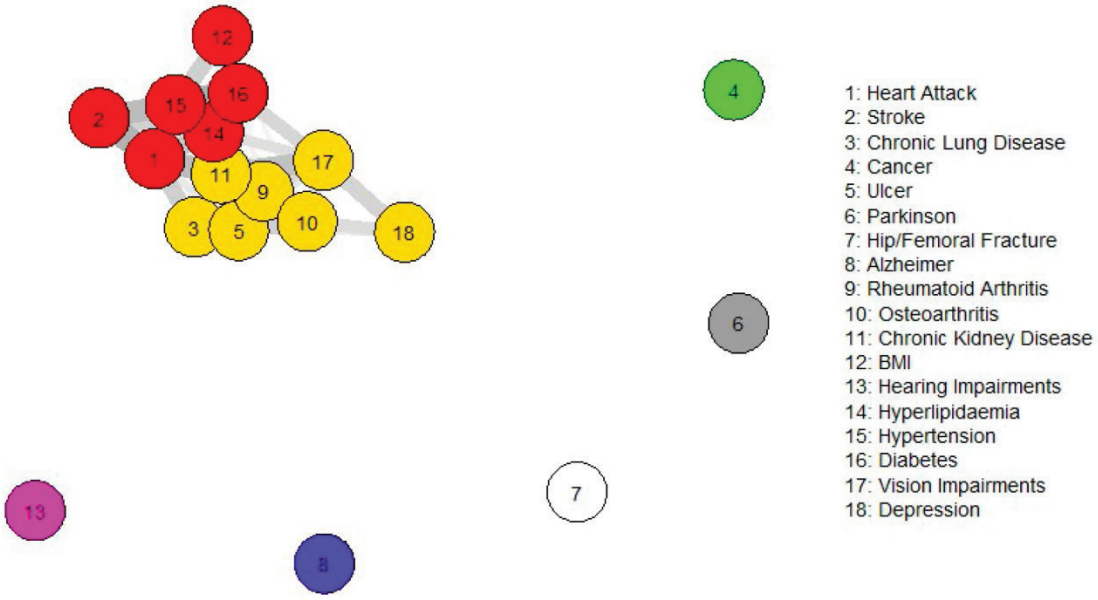


Figure 3. MGM for females aged 50-64: All diseases, SHARE Wave 7.

Figure 3 illustrates the results of the MGM method applied to all diseases in females aged 50–64. The distribution appears more heterogeneous compared to that observed in males. For this group, the three clustering methods yielded identical results, which are partially consistent with those obtained for the male cohort. The RED cluster is again associated with cardiovascular and metabolic diseases, whereas the YELLOW cluster encompasses certain chronic conditions. The remaining five clusters each correspond to a single disease, with no apparent relationships among them or with the two larger clusters.

The strongest relationships in this network are the ones between Heart Attack, Stroke, Hypertension and Hyperlipidaemia, and between Heart Attack, Chronic Kidney Disease and Vision Impairments. Heart attacks, chronic kidney disease, and vision impairments could be interrelated through common conditions such as diabetes and hypertension, systemic effects like inflammation and endothelial dysfunction, and the impacts of treatments and lifestyle factors (this part in yellow can be dropped).

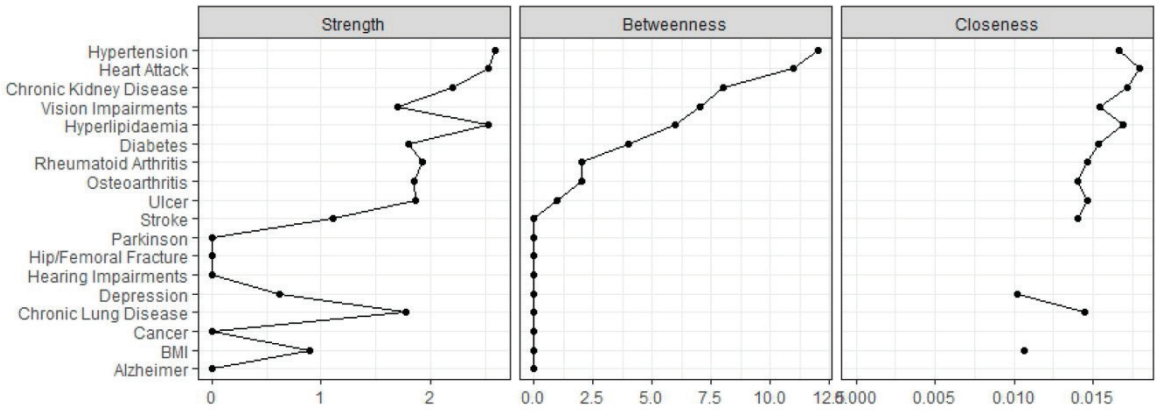


Figure 4. Centrality plot for females aged 50-64: All diseases, SHARE Wave 7.

Figure 4 shows that the nodes with the highest centrality strength are the ones related with

cardiovascular diseases: Hypertension, Heart Attack and Hyperlipidaemia. The nodes with the highest centrality betweenness, also in this case, are the ones corresponding to Hypertension and Heart Attack. The nodes with the highest centrality closeness are Heart Attack, Hyperlipidaemia, Hypertension and Chronic Kidney Diseases. It is possible to notice, also in this case, that there are nodes with no closeness. These nodes are Parkinson, Hip/Femoral Fracture, Hearing Impairments and Cancer. These nodes have no relationship with each other and with the other nodes.

In the context of a Mixed Graphical Model (MGM), a node having no links (or edges) means that the variable represented by that node does not have significant associations with other variables in the model. The variable represented by the node is conditionally independent of all other variables in the model given the observed data. This means that knowing the value of this variable does not provide additional information about the values of other variables when conditioned on the remaining variables.

Mixed graphical models often use regularization techniques, such as L1 regularization (Lasso), to enforce sparsity and avoid overfitting. Regularization can prune weaker connections, leaving some nodes without any links if their associations with other variables are not strong enough.

For subjects aged 65+, results not reported here, the clusters were fewer and better defined, with notable aggregations for cardiometabolic and for neurodegenerative diseases linked with dementia. Stroke was associated with neurodegenerative diseases in this age group for both genders.

4. Conclusions

In this paper mixed graphical models are considered to explore the complex interplay of comorbidities among patients using SHARE data. MGMs proved effective in modelling intricate dependencies between comorbidities, identifying interesting interconnections potentially driven by shared risk factors. Unlike other approach to study multimorbidity patterns such as factorial analysis or clustering techniques, the combination of graphical model and network analysis not only group chronic diseases into meaningful patterns for each age group, but also model the relationship between these diseases, adding predictive power and hence making the established network more informative.

Due to space constraints, we present only selected results of the gender-based analysis related to the 50–64 age group. However, results for the 65+ age group are available upon request.

The clustering algorithms Fast-greedy, Louvain, and Optimal produced consistent results across SHARE wave 7 and 9 datasets, showing a robust underlying structure in the data. Analysis by age group revealed distinct patterns of disease clustering, as expected. In this short paper, we have shown only some results. Among subjects aged 50-64, frequent clusters involved cardiometabolic and other chronic diseases.

Other results at our disposal have revealed that excluding diseases with a prevalence under 1% led to fewer clusters, emphasizing the significance of prevalent diseases in clustering outcomes. The study also considers the potential for built-in redundancy in networks, when removing a central node like, for example, hypertension may not drastically alter the overall network due to compensatory mechanisms in both waves for each possible gender and age class combination.

Some limitations should be reported, such as the exclusion of institutionalized individuals in nursing homes from the SHARE datasets, potentially leading to a healthier sample population, this exclusion may limit the generalizability of the findings.

We are confident that these approaches can bridge the gap between epidemiology, statistical modeling, and healthcare delivery by leveraging advanced analytical tools to unravel the complex web of multimorbidities prevalent in aging populations. The relational structure of diseases, as uncovered by MGM, can guide resource allocation in health systems, ensuring interventions target areas with the greatest potential impact on patient outcomes and system efficiency.

Acknowledgements

This research was funded by the 2020 PRIN-Research Project of National Relevance "Social and health Frailty as determinants of Inequality in Aging (SOFIA)".

We acknowledge co-funding from European Union – Next Generation EU, in the context of The National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 "Conseguenze e sfide dell'invecchiamento", Project Age-It (Ageing Well in an Ageing Society).

This paper uses data from SHARE Waves 7 and 9. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GAN°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1- AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

References

Bergmann, M., Wagner, M., Börsch-Supan, A. (Eds.) (2024). *SHARE Wave 9 Methodology: From the SHARE Corona Survey 2 to the SHARE Main Wave 9 Interview*. Munich: SHARE-ERIC.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S. (2013). Data resource profile: The survey of health, ageing and retirement in Europe (SHARE). *International Journal of Epidemiology* DOI: 10.1093/ije/dyt088.

Csardi, G., Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J. , Schmittmann,V. D., Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software* 48 (4), 1–18.

Ising, E. (1925) Beitrag zur theorie des ferromagnetismus", *Zeitchrift für Physik*, 31, no.1, pp. 253-258.

Yang, E., Baker,Y., Ravikumar, P., Allen, G., Liu, Z. (2014). Mixed graphical models via exponential families. In S. Kaski J. Corander (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, *Volume **33** of Proceedings of Machine Learning Research, Reykjavik, Iceland*, pp. 1042–1050. PMLR.

Haslbeck, J. M. B. , Waldorp, L. J. (2020). mgm: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software* **93** (8), 1–46.

SHARE-ERIC (2024). *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 9.0.0.* SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w7.900

SHARE-ERIC (2024). *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 9. Release version: 9.0.0.* SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w9.900.

Smith, N. R., Zivich, P. N., Frerichs, L. M., Moody, J., Aiello , A. E. (2020). A guide for choosing community detection algorithms in social network studies: The question alignment approach. *American Journal of Preventive Medicine* 59 (4), 597–605.

doi.org/10.26398/asaproc.0095

Analysing fan engagement and sentiment on social media following Bologna FC's 2023/2024 Champions League qualification

Giulia Cavrini [a], Agostino Stavolo [a,b]
[a] Faculty of Education, Free University of Bozen, Bozen, Italy
[b] Department of Social Sciences, University of Naples Federico II, Naples, Italy

1. Introduction

Sentiment analysis (SA) is a computational technique used to study human behaviour, perspectives, and feelings. This analysis can be applied to any subject, whether it is a person, event, or topic. As a key component of natural language processing (NLP), SA aims to extract and interpret opinions and attitudes from textual data (Birjali et al., 2021; Chaturvedi et al., 2018) highlighting positive and negative sentiments, and categorizing it as a text-classification task. With the overwhelming volume of textual data available and the impracticality of manual text evaluation, effective sentiment analysis techniques have become invaluable (Choi and Lee, 2017). While humans can effortlessly interpret opinions and attitudes in written language, replicating this capability through algorithms presents a significant challenge (Mäntylä et al., 2018). As a result, SA has emerged as one of the most rapidly expanding research areas in computer science, fuelled by the explosion of user-generated content on the web (Piryani et al., 2017).

The rapid expansion of social media has significantly enhanced public engagement with football events, enabling fans to transcend geographical boundaries and form communities. Sporting events evoke strong emotions among fans and thus act as natural laboratories to explore emotions and unfold them in real-time. This digital revolution has allowed football enthusiasts to actively contribute to content produced by professional media creators, making social media a crucial platform for sharing opinions and daily life events (Ljajić et al., 2015). Computational tools, such as sentiment analysis, provide new ways to examine these dynamic emotional processes. The immense popularity of football makes it an ideal subject for analysing fan sentiment and comments, particularly on platforms like Facebook.

This study focuses on the Bologna FC 1909 football team, especially following their historic qualification for the Champions League in 2023/2024 after a 60-year hiatus. By capturing fans' reactions and emotions to various in-game events, the data offers deep insights into the evolving dynamics of fan sentiment as matches progress. Using latent Dirichlet allocation (LDA) and SA, the study aims to uncover the main themes dominating discussions among Bologna FC fans and provides a comprehensive view of fan interactions during the recently concluded Serie A championship.

2. Background

SA has become a widely used automated process for examining a text's polarity, typically categorized into two main approaches: machine learning and rule based. The machine-learning approach involves sentiment models that have been previously trained on labelled textual datasets. This method requires significant time and effort for manual text labelling and is heavily dependent on the quality of the training data. On the other hand, the rule-based approach does not rely on human-labelled documents. Instead, it uses word lexicons, where each word or sequence of words is associated with a specific sentiment score.

The machine-learning approach is considered domain-dependent, meaning that models trained for one domain may not perform well in others. In contrast, the rule-based approach is often

favoured in the literature due to its simplicity and the relative ease of adapting sentiment lexicons for specific domains. The latter method has proven to be efficient and sufficiently accurate, making it increasingly popular for analysing social media data (Kumaresh, 2022).

Specifically, when dealing with opinions, it is particularly valuable to consider the semantic orientation of the texts, which involves expressing the sentiment in a numerical form (Cavique et al., 2022). In this study, we performed sentiment analysis with R package *sentimentR*. To calculate the polarity score for each document, the text is first divided into its constituent sentences to capture the sentiments expressed in the posts and comments of the football club supporters. For a given document d_i, denoted with i=[1,n], its q_i sentences $\{s_{i1}, ..., s_{ik}, ..., s_{iqi}\}$ are identified using strong punctuation marks such as full stops, question marks, and exclamation marks as separators. The k-th sentence is then represented as a sequence of terms $\{w_{ik1}, ..., w_{ikj}, ..., w_{ikpk}\}$, maintaining the term order within the sentence. Each term w_{ikj} in the *k-th* sentence of the *i-th* document is compared to a dictionary of polarized terms, assigning a score (PS_{wikj}) of -1 for negative terms and +1 for positive terms. Terms not present in the dictionary are considered neutral and assigned a score of 0. The polarity of each term is further 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 for the enhancement or attenuation of the negativity or positivity of each term, contributing to a more nuanced measure of polarity. The polarity score (PS_{sik}) for each sentence is obtained by summing the weighted term scores (PS_{wikj}) divided by the square root of the sentence length:

$$PS_{S_{ik}} = \frac{\sum_{j=1}^{Pk} PSw_{ikj}}{\sqrt{P_k}} \quad (1)$$

Since the goal is to derive a polarity score at the document level, an overall score (PS_{di}) is calculated for each text. This is achieved by taking a down-weighted zero average of sentence polarities, assigning a lesser weight to sentences conveying a neutral sentiment:

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

Here, q_i^{NEG}, q_i^{POS}, and q_i^{NEU} represent the number of sentences in d_i with negative, positive, or neutral polarity, respectively (Marino et al., 2022).

3. Methodology

We selected Facebook groups such as "Bologna FC 1909" and "Forza Bologna Sempre". We used Crowd Tangle software for extracting 708 posts and 30.770 comments. Since comments and posts are written as free text, the opinions they contain are encoded in a form that is difficult to process automatically.

Text analysis involves a series of essential steps for extracting meaningful information from a document collection. Initial processes include tokenization and parsing to break down the text into distinct tokens, reducing it to basic graphic forms. Normalization phases, such as converting characters to lowercase and grammatical tagging, follow to ensure consistency. Lemmatization is then applied to reduce words to their base forms, simplifying the analysis. Stop-words are removed to focus on meaningful terms, leading to the construction of a vocabulary with unique terms and relative occurrences. The final step involves creating a document-term matrix, where rows represent documents and columns represent terms.

4. Preliminary results

Table 1 reports the four topics extracted by LDA. The optimum number of topics is the shorted

distance between metrics, as a reverse ratio of dissimilarity (De Mascellis et al., 2022). The number of topics that both minimize the distances of Arun et al. (2010) and Cao et al. (2009) and maximize the distances of Griffiths and Steyvers (2004) and Deveaud et al. (2014) is $k = 4$.

Table 1 – Topics extracted from LDA

Social initiatives	Fanbase	Game Dynamics	Awards
Donazione	Tifoso	Goal	Prestito
Donare	Presidente	Giocare	Amicizia
Sinisa	Thiago_Motta	Giocatore	Orso
Cancro	Cuore	Vincere	Dispiacere
Diretta	Sogno	Perdere	Giornata
Malattia	Emozione	Campionato	Premio
Aiutare	Biglietto	Rigore	Vincitore
Attività	Champions_League	Allenatore	Stadio
Figli	Maglia	Campo	Interesse
Moglie	Qualificazione	Arbitro	Pregare

The first theme focuses on social and humanitarian initiatives, particularly linked to prominent figures like Sinisa Mihajlovic and battles against serious illnesses. The community's active participation in these social causes, illustrated by their involvement in events and direct support, reflects the importance of solidarity campaigns among Bologna FC fans. This creates a strong sense of unity and social responsibility within the community.

The second theme revolves around the passion and devotion of the fans for football. It underscores the intense loyalty and love for the team and its leaders, manifesting through deep emotional engagement. Fans' emotional investment in matches and events related to the team is evident, signifying their unwavering support and participation throughout the championship.

The third theme delves into the performance and dynamics of the game itself. It points to a keen focus on the technical and competitive aspects of football, highlighting how fans closely follow matches and analyse the strategies and decisions impacting the team's performance. This theme showcases the critical and passionate engagement of fans with the game's intricacies.

The fourth theme explores social relationships and recognitions within the football context. It emphasizes the importance of personal interactions and behind-the-scenes stories, valuing the recognition and shared experiences in the sports environment. This theme illustrates how fans appreciate not only the athletic performances but also the personal stories and individual successes, both on and off the field.

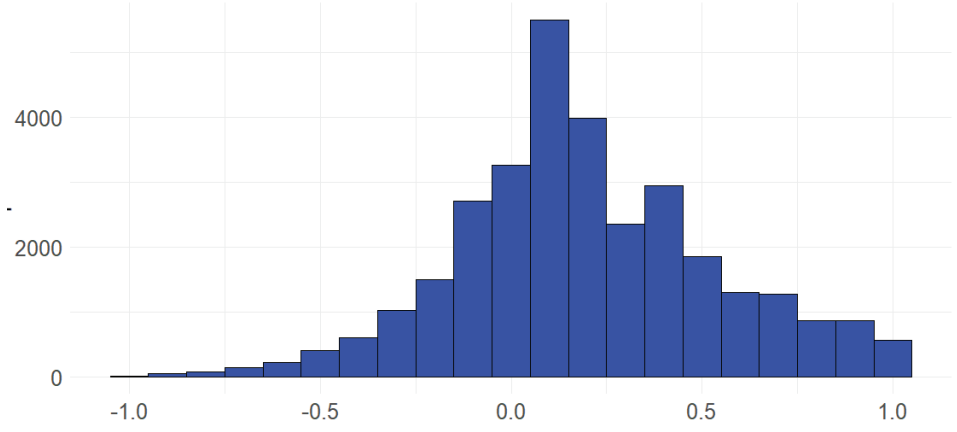


Figure 1 – Sentiment distribution

Figure 1 illustrates the distribution of sentiments expressed by Bologna FC fans, ranging from -1 (very negative) to 1 (very positive). The distribution shows a pronounced peak around 0, representing the point of maximum frequency, with frequencies gradually decreasing towards the extremes of very negative (-1) and very positive (1) sentiments. The sentiment distribution is symmetrical, with a slight inclination towards the positive side. This suggests that fans' reactions are generally balanced, with a slight bias towards the positive. This balance likely reflects the team's performance, with an even mix of wins and losses over the season.

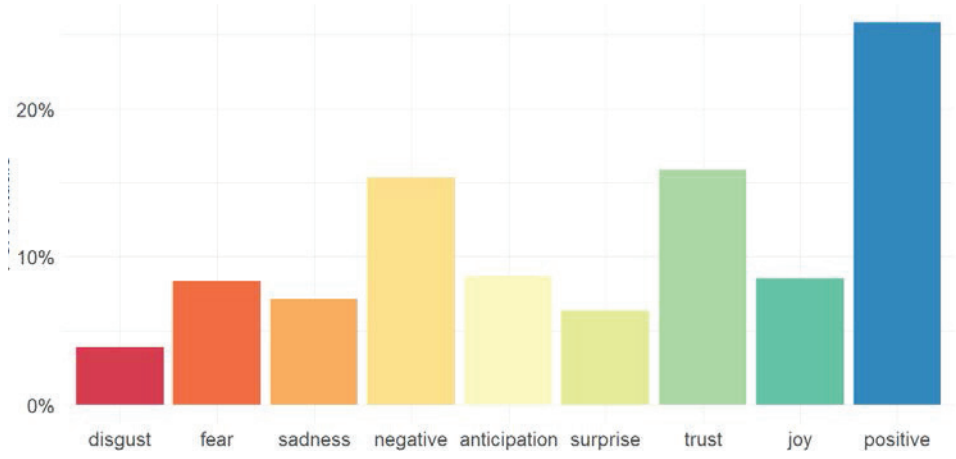


Figure 2 – Emotional analysis

In Figure 2 it is evident that the 'positive' sentiment is the most widespread, accounting for over 20% of the total, reflecting moments of victory or particularly positive performances of the team. This is followed by the "trust" sentiment ("*non lo conoscevo ma sembra che i numeri li abbia e ho fiducia nella capacità di Sartori di scovare talenti a buon prezzo -"I didn't know him, but it seems he has the numbers, and I trust Sartori's ability to find talents at a good price"*), which indicates a stable level of trust in the team and is linked to the choices the coach has made in the purchase of players, and the 'negative' sentiment, which arises in response to defeats or disappointing performances. Feelings such as 'disgust' ("*Il rigore non c'era: l'arbitro ha dato il vantaggio, poi il Milan ha perso il possesso e lui ha fischiato. Rigore inventato e Motta espulso! Una decisione scandalosa* – "*The penalty wasn't there: the referee played the advantage, then Milan lost possession, and he blew the whistle. A fabricated penalty, and Motta was sent off! A scandalous decision"*), 'fear' ("*mi fanno più paura certi arbitraggi*" - *"I'm more afraid of certain refereeing decisions.")* and 'surprise' ("*allo stadio tutti a chiedere a gran voce uomini freschi per il finale ma Thiago sorprende sempre e fidiamoci che ne sa più di noi*" – *"At the stadium, everyone was loudly calling for fresh players for the final minutes, but Thiago always surprises us. Let's trust him; he knows more than we do.")* are less common but occur mainly in matches misjudged by the fans. Finally, feelings such as 'joy' and 'anticipation' have similar percentages, around 10%, while 'sadness' is present but less frequent, linked to moments of defeat or injuries to key players.

Figure 3 shows the trend in the average sentiment expressed by Bologna FC fans over time, from August 2023 to June 2024. A notable peak in average sentiment is observed in October 2023. This is attributed to the fans' enthusiasm for the start of the championship and the victories achieved by the club, as well as good expectations on the purchases made during the football market period. Thereafter, there is a significant drop from November 2023 to January 2024, with the lowest point in January 2024, in line with the club's football results, which sees a series of defeats and draws that are not appreciated by the fans, criticising the coach for the choices made and the lack of an optimal playing strategy. After January, average sentiment shows an upward trend, with a noticeable increase from May 2024 and a peak in June 2024, suggesting a particularly positive end to the season, linked to the team's qualification for the Champions League.

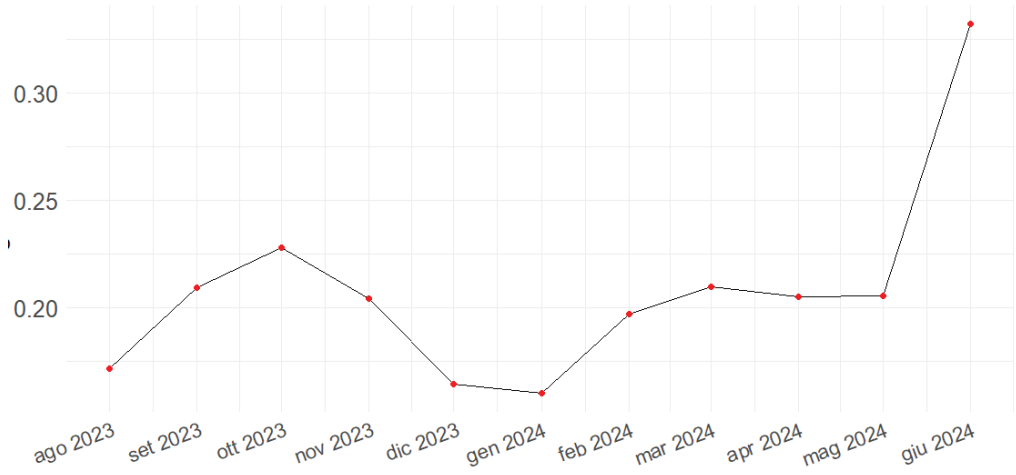


Figure 3 – Average sentiment over time

4. Conclusion

This work aimed to provide an initial exploratory contribution on the topic. We have seen that following the success of the Bologna FC football team and their subsequent qualification for the Champions League after 60 years, most users on Facebook expressed a positive sentiment, characterized by emotions of joy, happiness, and trust, but also fear of not being able to achieve the set goals. It has been shown that during the football season, the sentiment distribution reached its peak during the last matches (May-June) at the time of qualification, while the winter phase proved to be more challenging and not always well received by the fans. The analysis demonstrates how social media platforms serve as a crucial space for fans to express their emotions, engage in discussions, and form communities. The insights gained from this analysis can help sports organizations and marketers to better understand fan behaviour, enhancing engagement strategies and fostering a stronger connection with their audience.

The work faces several limitations, primarily due to the reliance on the Italian language for sentiment analysis. This dependency necessitates the creation of a context-specific dictionary tailored to both the language and the domain of sports events. Sentiment analysis in Italian presents unique challenges that may not be as prevalent in more commonly studied languages, such as English. Italian has various linguistic nuances, idiomatic expressions, and regional dialects that can complicate the accurate interpretation of sentiments. Additionally, the need to develop a specialized dictionary for sports events adds another layer of complexity. Commonly used sentiment analysis tools and lexicons are often not directly applicable to the sports domain, requiring significant customization.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., Narasimha Murthy, M.N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conferenc*e, PAKDD 2010, Hyderabad, India, Proceedings. Part I 14 (pp. 391-402). Springer Berlin Heidelberg.

Birjali, M., Kasri, M., Beni-Hssane, A.A. (2021). Comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, **226**: 107134.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9), pp. 1775-1781.

Cavique, M., Ribeiro, R., Batista, F., Correia, A. (2022). Examining Airbnb guest satisfaction tendencies: A text mining approach. *Current Issues in Tourism*, **25**(22), pp. 3607-3622.

Chaturvedi, I., Cambria, E., Welsch, R. E., Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, **44**, pp. 65-77.

Choi, Y., Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers*, **19**, pp. 993-1012.

De Mascellis, A.M., Grassia, M.G., Stavolo, A. (2022). A. topic modeling and sentiment analysis: An analysis of tweets about the game cyberpunk 2077. *In JADT 2022: 16th International Conference on Statistical Analysis of Textual Data*, pp. 282-288.

Deveaud, R., SanJuan, E., Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), pp. 61-84.

Griffiths, T.L., Steyvers, M. (2004). Finding scientific topics. *PNAS,* 101(1), pp. 5228-5235.

Kumaresh, S. (2022). Sentiment analysis of COVID-19 vaccine in a social media platform using machine learning techniques. *Syndicate-The Journal of Management*, **39**.

Ljajić, A., Ljajić, E., Spalević, P., Arsić, B., Vučković, D. (2015). Sentiment analysis of textual comments in field of sport. *In 24th International Electrotechnical and Computer Science Conference* (ERK 2015), IEEE, Slovenia.

Mäntylä, M. V., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, **27**, pp. 16-32.

Marino, M., Mazza, R., Misuraca, M., Stavolo, A. (2022). Monitoring consumer sentiment using control charts. In *Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment*. pp. 108-113.

Piryani, R., Madhavi, D., Singh, V.K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, **53**(1), pp. 122-150.

doi.org/10.26398/asaproc.0096

Text mining of tourist reviews: Sentiment analysis and trends in the Apulian tourism

Dario Centorame[a], Giovanni Pino[a], Annalina Sarra[b] , Eugenia Nissi[a]
[a] Department of Economics, University of Studies 'G. d'Annunzio', Chieti–Pescara, Italy
[b] Department of Socio-Economic, Management and Statistical Studies, University of Studies 'G. d'Annunzio', Chieti-Pescara, Italy

1. Introduction

In recent years, there has been an exponential increase of digitalisation processes across various domains. One of the sectors experiencing the most dramatic changes is tourism, where communication and marketing practices are experiencing profound changes to adapt to the demands of a more digitalized world. In this context, the emergence of online tourism platforms such as Yelp, Google Maps, TripAdvisor, and Airbnb offer tourists the chance to explore locations and attractions autonomously. Due to the diffusion of similar platforms, user generated content (UGC) has become increasingly important, both for tourists and destination marketing organizations -DMOs- (Lam et al., 2020; Assaker, 2020) and now guides decision-making processes (Egger et al., 2022; Guo et al., 2017).

However, the analysis and integration of UGC challenges traditional qualitative methods. Given the 'big data' nature of these datasets—characterized by volume, velocity, variety, and value—researchers and practitioners seek to address these challenges using quantitative methods, ideally with machine learning approaches due to their ability to solve complex and data-intensive tasks (Bishop, 2006; Goodfellow et al., 2016).

In this study, we focus on TripAdvisor, a prominent platform widely utilized by tourists for reviewing and sharing information about their travel experiences. We content-analysed 74,721 reviews of various attractions located in Apulia, a southern Italian region. To gain a deeper understanding of tourists' perspectives, we applied sentiment analysis and topic modeling techniques.

The paper is structured as follows. Section 2 provides an overview of the data collection process and the methodological approach adopted to analyse the dataset. Section 3 presents the key results of our analysis. Finally, the concluding section synthesizes the findings and offers a comprehensive summary of the main insights emerged from the study.

2. Data and methods

2.1 Tourists' reviews collection

We concentrated on the "Things to Do" section of TripAdvisor, which covers various attractions. For each of the 64 cities listed in Apulian register of tourist-oriented municipalities[1],we selected up to 5 attractions per city, considering the ones with the highest number of reviews. We then ran a scraping algorithm using the Python programming language. We limited this process to text written in Italian, by far the most frequent language, and obtained a total of 74,721 reviews regarding the period 2010 - June 2024. The dataset includes detailed information such as the visit period (month, year), location, username, type of visit, review text, attraction rating on a 5-point scale, and upvotes received from other community members. More specifically, the reviews are primarily concentrated in recent years with a peak in 2017 (13,850 reviews). The number of reviews per year declined during the COVID-19 pandemic (2020-2022) but increased slightly in the following years. As noted in Xiang et al. (2017), reviews from platforms such as TripAdvisor are predominantly positive. Similarly, we

[1] https://www.regione.puglia.it/web/turismo/elenco-comuni-ad-economia-turistica-e-citta-d-arte

found 46,809 reviews rated 5 stars and 19,370 rated 4 stars, with fewer than 4% of reviews rated 1 or 2 stars. Additional information is shown in Table 1 below.

We pre-processed the reviews to prepare the text for the analysis. Specifically, we tokenized the sentences, removed punctuation marks, converted the reviews' text to lowercase, and eliminated common Italian stop-words and emoticons. After this preprocessing stage, only 116,748 unique tokens remained from an initial pool of 203,673 and each review included an average of 27 words.

Table 1: Distribution of reviews by year and city

Years	Reviews
2010	3
2011	121
2012	1.117
2013	2.170
2014	4.045
2015	8.356
2016	13.266
2017	13.850
2018	11.774
2019	8.727
2020	4.702
2021	2.644
2022	2.035
2023	1.560
2024	351

City	Reviews
Lecce	6.899
Bari	5.818
Taranto	5.620
Polignano a Mare	3.756
Alberobello	3.555
Andria	3.189
Trani	3.119
Otranto	2 837
Porto Cesareo	2.500
San Giovanni Rotondo	2.500
Castellana	2.415
Gallipoli	2.218
Peschici	1.920
Castrignano del Capo	1.810
Others	26.579

2.2 Sentiment Analysis

Sentiment Analysis (SA) (Alaei et al., 2019; Bi et al., 2024) is a valuable tool to examine TripAdvisor reviews. SA uses natural language processing (NLP) techniques to identify and extract subjective information from text, typically expressed with a positive, negative, or neutral sentiment.

To implement SA, we employed a combination of lexicon-based and machine learning-based approaches. The lexicon-based methods involved using predefined dictionaries of words associated with specific sentiments. In particular, we relied on two well-known and widely used lexicons: *OpeNER,* developed by the Italian National Research Council, and the *NRC Emotion-Lexicon* from the National Research Council of Canada. The *OpeNER* lexicon is capable of distinguishing sentences into positive, negative, and neutral categories, while the *NRC lexicon* recognizes not only positive and negative sentiments but also eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), providing richer insights.

For the machine learning approaches, we used two pretrained transformer models freely available in the *Huggingface* repository: *Feel-IT*, which infers sentiments using the Italian language, and *Emotion English DistilRoBERTa-base*[2]. The latter model was utilized by first translating the sentences from Italian to English using the *MarianMT*[3] framework and the *Helsinki-NLP/opus-mt-it-en*[4] model.

2.3 Topic modeling

We then implemented the latent Dirichlet allocation (LDA) technique, one of the most popular methods for extracting latent topics from textual data in an unsupervised manner, which is widely used in tourism research (Kirilenko, 2021). As part of the broad family of probabilistic topic models, LDA is a generative model of a corpus (Blei et al., 2003) that assumes the existence of a hidden

[2] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base
[3] https://huggingface.co/docs/transformers/en/model_doc/marian
[4] https://huggingface.co/Helsinki-NLP/opus-mt-it-en

structure in a text, spanning multiple dimensions (i.e., topics), each of which is characterized by a specific distribution over words.

Following the preprocessing phase described earlier, where we tokenized and normalized the text, the tourists' reviews were converted according to the bag of words model. In this representation, each word is indexed using a fixed-size vocabulary, and each review is vectorised based on these indices and the frequency of words. With this transformed document representation, the LDA model is able to infer latent topics hierarchically, based on the co-occurrence of words within the reviews.

In addition to LDA, we explored dynamic topic models (DTM) (Blei and Lafferty, 2006), a technique that extends LDA by incorporating a time dimension. DTM are probabilistic time series models specifically designed to analyse the evolution of topics over time. This extension allows us to track changes in topics across different years and understand how tourists' posed on SA. TripAdvisor allows users to rate attractions on a star rating system, which can orientate prospect visitors' decisions. However, recent research suggests that this star rating system can be overly simplistic and may not capture the full depth of users' experiences. Therefore, to fully explore the content of the reviews, we employed various sentiment analysis techniques as described in the previous section.

To determine the most effective method for further analysis, we used a framework developed by Kirilenko et al. (2021) and Taecharungroj (2019), which classifies reviews as "positive" if the rating is equal to or above four stars and "negative" if it is lower than four stars. We evaluated the performance of our sentiment analysis methods using standard classification metrics, including Accuracy, Precision, Recall, and F1-score (which combines Precision and Recall). The results of these evaluations are opinions and patterns of reactions have evolved over time due to shifts in sentiments and interests.

3. Main results

The analysis of TripAdvisor reviews was conducted using the Python programming language and standard packages used for NLP tasks, such as *NLTK* for pre-processing, *Scikit-Learn, Huggingface Transformers pipeline* and *Gensim* for machine learning modeling.

3.1. Sentiment Analysis

We initially focused on SA. TripAdvisor allows users to rate attractions on a star rating system, which orientates prospect visitors' decisions. However, recent research suggests that this star rating system can be overly simplistic and may not capture the full depth of users' experiences. Therefore, to fully leverage the content of the reviews, we employed various sentiment analysis techniques. Our approach involved a combination of lexicon-based methods and machine learning models. To determine the most effective method for further analysis, we used a framework from Kirilenko et al. (2021) and Taecharungroj (2019) that classifies reviews as 'positive' if the rating is equal to or above four stars and 'negative' if below four stars. We evaluated the performance of our sentiment analysis methods using standard classification metrics, including accuracy, precision, recall, and F1-scores. The results of these evaluations are summarized in Table 2.

Table 2: Performance metrics for sentiment analysis techniques

Technique	Accuracy	Precision	Recall	F1-score
NRC emoticon-lexicon	0.8272	0.8857	0.9240	0.9045
CNR lexicon	0.8842	0.9187	0.8842	0.9008
DistilRoBERTa-base	0.7322	0.8852	0.8793	0.8823
Fell-IT	0.7688	0.8845	0.8498	0.8668

The results indicate that the NRC emoticon-lexicon is the most effective sentiment analysis technique, achieving the highest F1-score (0.9045) and a strong recall (0.9240), which suggests it

adequately captures and classifies sentiments. Despite its slightly lower accuracy (0.8272), its overall performance surpasses the other models in balancing precision and recall.

3.2. Topic modeling

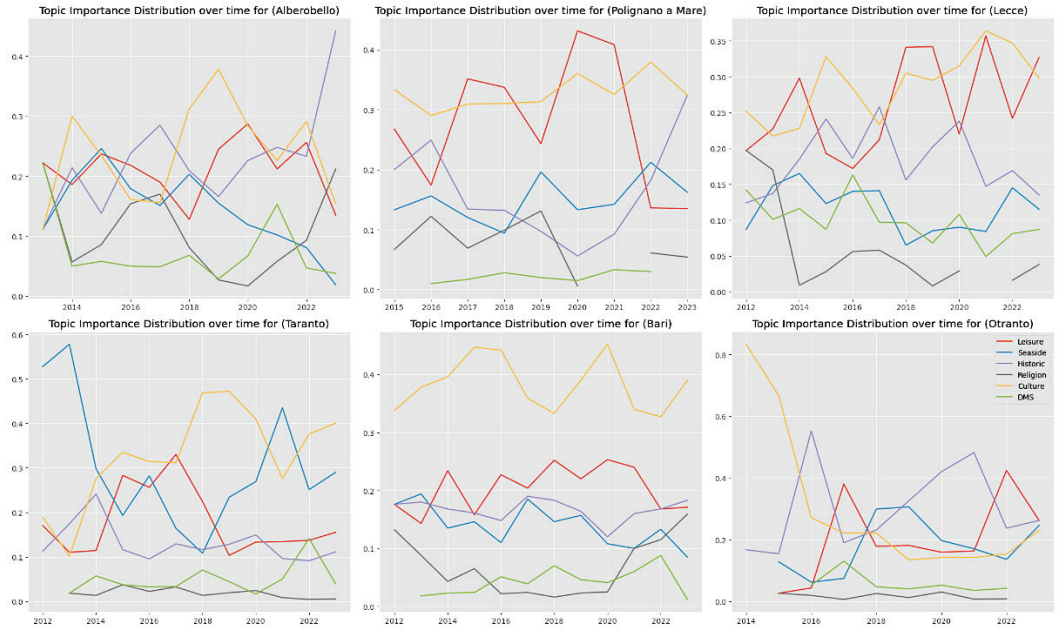
To uncover latent themes within the reviews, we implemented topic modeling. Toward this end, we applied LDA with TF-IDF correction to refine the data, using the hyperparameters α and β as suggested by Tang et al. (2014) and Griffiths and Steyvers (2004). α, set to approximately 50/T, where T denotes the number of topics, α controls topic distribution per document, while a small β (0.1) encourages the model to use a wide range of words for each topic, rather than concentrating words in a few topics. The static LDA analysis revealed six key topics reflecting various aspects of tourist experiences: *Destination Management and Services*, *Seaside and Urban Amenities*; *Historic, Artistic, and Cultural Attractions*; *Leisure Opportunities*; *Natural Amenities*; and *Religious Attractions*. Each topic is characterized by keywords that highlight specific themes discussed in the reviews, providing insights into tourist preferences and experiences.

For the dynamic modeling phase, we utilized *Gensim*'s *ldaseqmodel* to analyse the evolution of topics over time. This method processes the corpus in time-sliced segments of reviews, fitting LDA models iteratively so that the parameters for each time slice remain consistent with those of the previous slice. Specifically, the model adjusts the α and β parameters across contiguous time slices using the formula:

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1.k}, \sigma^2 I) \text{ and } a_t|\, a_{t-1} \sim N(a_{t-1}, \delta^2 I)$$

Our dynamic analysis excluded the years 2010 and 2011 due to limited sample sizes and the year 2024 due to incomplete data. We maintained six predefined topics for consistency with our static analysis. The dynamic analysis yielded topics similar to those identified previously, but due to slight variations and evolving trends, we renamed them as *Leisure*, *Seaside*, *History*, *Religion*, *Culture*, and *Destination Management and Services*. The results of this analysis for the main tourist cities of the region (Lecce, Bari, Taranto, Polignano a Mare, Alberobello, and Otranto) are presented in Figure 1.

Figure 1: Topic importance distribution over time



The results show that the data are affected by noise and irregularities, and there does not appear to be a consistent pattern across all cities. As regards the six most visited cities (Figure 1), we noticed

that, for Bari, the largest city with diverse tourism offerings, there is no noticeable evolution in review topics. In contrast, the results concerning Lecce, the second-largest city, show an overall increase in the frequency of the Leisure and the Culture topics, while the frequency of Destination Management and Services and the Religion topics has decreased.

We also noticed that, for Polignano a Mare, a city known for its seaside attractions, the Seaside theme is not the most prevalent across all years. Instead, Culture and Leisure are more dominant, with Culture maintaining a steady presence. Similarly, in Alberobello there is a noticeable rise in the frequency of the History topic, which is linked to an increase in the Religion topic in recent years. For Taranto, the Culture topic has shown an upward trend throughout most of the analysed period. The Seaside theme, after a significant drop in 2013-2014, has seen a resurgence in recent years For Otranto, there seems to be no significant evolution in the trends of the topic

By cross-referencing the results from the topic analysis with those from the sentiment analysis carried out with the NRC emoticon-lexicon, we assessed how the evaluation of the topics has changed over the years. Overall, there is a general positive trend across all topics between 2014 and 2020. However, as we know, the years 2020-2022 were marked by the global Covid-19 pandemic, which severely affected the tourism industry. This emerges from tourists' opinions: for all the topics, the average sentiments showed a downturn, which was particularly pronounced for the religion and history topics as we can see from Figure 2.

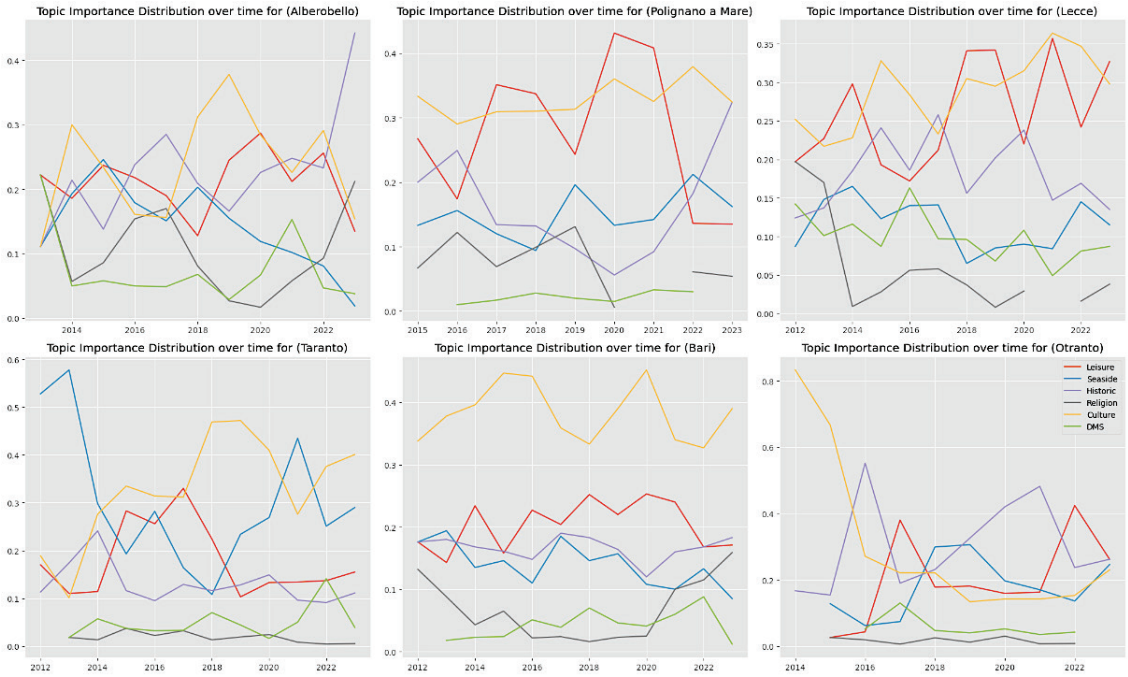


Figure 2: Trends in Topic Sentiment over Time

4. Conclusions

This study offers a detailed analysis of TripAdvisor reviews regarding Apulian attractions in southern Italy. It employed sentiment analysis and topic modeling to track changes in tourist opinions over a decade. The topic modeling identified six major themes: leisure, seaside, history, religion, culture, and destination management and services. These themes evolved over time, reflecting shifts in tourist interests and the influence of external factors like the pandemic. The sentiment analysis showed a generally positive trend between 2014 and 2020, indicating an increase in tourist satisfaction during this period. However, the COVID-19 pandemic (2020-2022) caused a significant decline in sentiment, especially for religion and history topics, highlighting the negative impact of this global emergency on cultural tourism.

This research contributes to the growing literature on tourism digitalization, with practical implications for destination management and marketing. Our approach emphasises the importance of combining sentiment analysis with topic modeling to gain deeper insights into tourist behaviour and offer valuable information to destination marketing organizations and other tourism stakeholders. Future research could explore the impact of specific events or policies on tourist sentiments and apply our approach to other regions, with the goal of supporting resilience and sustainability in tourism.

References:

Alaei, A.R., Becken, S., Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research,* **58**(2), pp. 175-191.

Assaker, G. (2020). Age and gender differences in online travel reviews and user-generated-content (UGC) adoption: Extending the technology acceptance model (TAM) with credibility theory. *Journal of Hospitality Marketing & Management,* **29**(4), pp. 428-449.

Bi, J.W., Zhu, X.E., Han, T.Y. (2024). Text analysis in tourism and hospitality: A comprehensive review. *Journal of Travel Research, 00472875241247318*.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* **3**, pp. 993-1022.

Blei, D.M., Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113-120), eds. W. Cohen, A. Moore, Pittsburgh (USA), pp. 113-120.

Egger, R., Gumus, O., Kaiumova, E., Mükisch, R., Surkic, V. (2022). Destination image of DMO and UGC on Instagram: A machine-learning approach, in *Information and Communication Technologies in Tourism 2022: Proceedings of the ENTER 2022 eTourism Conference*, eds. J.L. Stienmetz, B. Ferrer-Rosell, D. Massimo (pp. 343-355). Springer International Publishing.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.

Griffiths, T.L., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences,* **101**(suppl_1), pp. 5228-5235.

Guo, Y., Barnes, S.J., Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. *Tourism Management,* **59**, pp.467-483.

Kirilenko, A.P., Stepchenkova, S.O., Dai, X. (2021). Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply?. *Tourism Management,* **83**, 104241.

Lam, J.M., Ismail, H., Lee, S. (2020). From desktop to destination: User-generated content platforms, co-created online experiences, destination image and satisfaction. *Journal of Destination Marketing & Management,* **18**, 100490.

Taecharungroj, V., Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, **75**, pp. 550-568.

Tang, J., Meng, Z., Nguyen, X., Mei, Q. Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of the 31st International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* **32**(1), pp. 190-198

Xiang, Z., Du, Q., Ma, Y., Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, ***58***, pp. 51-65.

doi.org/10.26398/asaproc.0097

An exploratory analysis of the comments from the Istat's household budget survey questionnaires for designing an App*

Elena Cezza, Gabriella D'Ambrosio, Loredana De Gaetano, Edoardo Raimondi
Italian National Institute of Statistics, Rome, Italy.

1. Introduction

The Istat's household budget survey plays a fundamental role in understanding Italian families' spending habits. Indeed, by focusing on the consumption expenditure behaviours of households residing in Italy, this survey analyses the evolution and composition of household expenditures according to their main social, economic, and territorial features (Istat, 2016). In addition, it provides crucial data for implementing political decisions and economic strategies since it represents the informative basis for the official estimates of relative and absolute poverty in Italy and the measure of inflation by household expenditure classes (Istat, 2023).

Traditionally, the collection of data takes place through questionnaires in CAPI (Computer Assisted Personal Interviewing) technique and through the so-called "Diary", a paper form that is filled in independently by the respondents for a total period of 14 days. However, in the era of the digitalization of statistical processes, Istat is evaluating the possibility of also managing this Diary in digital format, by adopting a website or an application for mobile devices (App).

In this respect, the importance of designing a "digital Diary" also emerges from the comments provided by both the respondents and the interviewers, according to which the transition to an App could simplify the analysis and reduce the respondent burden associated with this statistical survey. This result is derived from a study on note fields of the Diary and the initial and final interviews recorded in the 2022 and 2023 survey editions. Overall, the present research takes into account 11,403 notes; however, not all of them have been analysed for the purposes of the present study since they did not refer to the difficulty in filling in the Diary/questionnaires; so, all the material examined refers to 4,800 notes (42,1% of the total).

Therefore, the present work aims to examine these users' comments in order to develop an application that integrates their usability needs with those of Istat for data accuracy and completeness, adopting a bottom-up approach.

2. The analysis of the comments provided during the initial interview

In the 2022 and 2023 survey editions, 2,813 comments (1,477 in 2022 and 1,336 in 2023) were recorded at the end of the initial CAPI interview. Among these, 157 were deemed as particularly relevant for the study. These comments were analysed using a mixed-method approach combining lexicometric and interpretative analysis (D'Amen and Pontecorvo, 2012)[1]. The lexicometric analysis, a quantitative method using statistical techniques and computational tools to identify patterns and frequency in word usage, was conducted using TaLTaC[2] software[2]. After pre-processing steps like parsing, normalization, and lemmatization, to prepare the corpus for analysis[3], lexicometric indexes were measured to assess the lexical richness of the comments. The type/token

(*) This paper is the result of the coordinated work of all the authors. However, Section 2 is attributed to Elena Cezza, Section 3 to Gabriella D'Ambrosio, Section 4 to Edoardo Raimondi, and Section 5 to Loredana De Gaetano. Section 1 is attributed to all the authors.

[1] Due to the limited number of observations and the similarity in themes, the notes from the two survey editions have been analysed collectively.

[2] TaLTaC[2] Version 2.11.2, https://www.taltac.com/download/

[3] For a definition of parsing, normalization, and lemmatization, see Giuliano and La Rocca (2012).

ratio, exceeding 20%, indicated insufficient richness for quantitative analysis (Giuliano and La Rocca, 2012). Therefore, the lexicometric approach was solely employed to extract the most frequent repeated segments[4], identified by the highest relative Significance Index[5], and use them as a framework for the subsequent interpretative analysis. This second phase, assisted by NVivo software[6], involved an open coding procedure, where the underlying meanings of the comments were interpreted and systematically categorized into themes (D'Amen and Pontecorvo, 2012). Nine codes were identified to summarize the content of the notes, which were then sorted into three macro groups, each representing a broader thematic domain (Table 1):

- *difficulties in filling in the Diary/questionnaire*: this category includes respondents who experience difficulties when addressing interview questions or foresee challenges in filling in the Diary, due to the complexity of the questions, limited language comprehension, or health/age-related issues;
- *completion method*: this group gathers notes on respondents who express a willingness to collect receipts and seek assistance from the interviewer in filling in the Diary;
- *willingness to participate in the survey*: this category relates to respondents who demonstrate reluctance to participate in the survey due to time constraints, concerns about potential fraud or privacy breaches, limited interest in statistical surveys, or other unspecified reasons. Additionally, this group comprises instances of respondents who are difficult to reach.

Table 1. Macro-categories and related codes identified for the initial interview notes

Macro-category	Code	Notes	
		N.	%
Difficulties in filling in the Diary/questionnaire	Language comprehension	46	29.3
	Complexity of the Diary/interview	19	12.1
	Health issues	9	5.8
Completion method	Receipts collection	6	3.8
Willingness to participate in the survey	Distrust and concerns over privacy	43	27.4
	Poor availability (reason not provided)	25	15.9
	Time constraints	6	3.8
	Respondent difficult to contact	3	1.9
Total		157	100.0

Source: Authors' elaboration on Istat data

3. The analysis of the comments provided in the Diary

The analysis of comments made by the respondents within the Diary has been carried out following an analytical approach and trying to identify themes, patterns, and core meanings (Patton, 2002), organize the textual material, and classify it into systematic categories (Neuendorf, 2002). Overall, the collected material refers to 5,473 comments included in the Diary (2,901 for 2022 and 2,572 for 2023), even though, after cleaning up the data matrix, not all the available material was useful for the analysis; consequently, there were 4,572 valid comments.

An initial analysis concerned the management of the Diary and the results of its compilation. As it is possible to note from Table 2, the Diary is partially filled in for both reference years (50.6% in 2022 and 51.3% in 2023). Then, there is 29.8% (2022) and 33.5% (2023) of those who did not fill the Diary at all while the remaining 19.6% and 15.2% who filled in the Diary, respectively for

[4] In TaLTaC2 terminology, repeated segments refer to sequences of adjacent words bound by two strong separators. For this study, the parameters for extraction were defined as follows: a segment length of four words, a minimum word frequency of two occurrences, and a minimum segment frequency of one occurrence.

[5] For a definition of the absolute and relative Significance Index, see Giuliano and La Rocca (2012).

[6] NVivo Version 14, https://lumivero.com/products/nvivo/

2022 and 2023, affirm they had faced several difficulties.

Table 2. Rates of Diary filled in during 2022 and 2023 survey editions

Outcomes	2022		2023	
	N.	%	N.	%
Diary filled in with difficulty	486	19.6	318	15.2
Diary partially filled in	1,255	50.6	1,074	51.3
Diary not filled in	739	29.8	700	33.5
Total	2,480	100.0	2,092	100.0

Source: Authors' elaboration on Istat data

More in detail, attempting to deepen the reasoning underlying the choices to fill in the Diary or not, the results of the analysis emphasize that:

- when the Diary is filled in with difficulty, the majority of respondents complain about the complexity of the survey. In this regard, the greatest difficulties concern the correct placing of expense items within the Diary, and the understanding of both tables on payment instruments and on places by purchase. Furthermore, the respondents claim that the Diary is a too long and dispersive paper form, too complex and detailed to be filled in, which requires particular attention and a lot of time to be dedicated to it. These difficulties are more marked in the case in which elderly people or foreign citizens with linguistic difficulties have to record the expense information;
- when the Diary is only partially filled in, the reasons refer more to the habits of the daily life of the individual/family interviewed. By way of example, it is possible to note the reduction of expenses due to the presence of food supplies, the absence from home for holidays or work reasons, food production, and the high frequency of lunches/dinners spent outside the home with relatives and/or friends. Last but not least, the impossibility of filling in the Diary for the entire reference period due to health problems, sometimes forcing the respondent to hospitalization;
- when the Diary is not filled in at all, it is possible to identify the reasons into three distinct categories: the difficulty in placing the expense items into the Diary; the loss of receipts/Diary; personal reasons (such as family losses). In the first of these three cases, although the Diary appears empty when it is withdrawn during the final interview, it is however possible to reconstruct the respondent's purchasing habits thanks to the collection of the receipts (vice versa, this is not possible in case of the loss of the receipts).

For a brief summary of the above, see Table 3.

Table 3. Main reasons underlying the difficulty, the partial and the missing filling in of the Diary

Outcomes	Main Reasons
Diary filled in with difficulty	• Difficulty in placing the expense items • Length and complexity of the Diary
Diary partially filled in	• Aspects of Daily life of respondents (i.e., presence of food supplies, absence from home for holidays or work reasons, food-self production, etc.)
Diary not filled in	• Difficulty in placing the expense items • Loss of receipts/Diary • Personal reasons

Source: Authors' elaboration on Istat data

4. The analysis of the final interview and key features for a digital Diary

The analysis concerning the note fields of the final interview involved 1,645 comments related to the year 2022 and 1,472 related to the year 2023, for a total of 3,117 notes. An analytical content analysis, as described in the third paragraph, was conducted on these, at the end of which 71 notes were evaluated as 'relevant' according to the following steps: *a*) identification of notes with relevant content; *b*) classification of comments; and *c*) identification of possible solutions. Subsequently, four categories of critical issues emerged, structured as follows:

- difficulties in contact, which includes cases where the interviewer is unable to contact the family because they are absent, distrustful, or not very participative;
- difficulties in completion, which comprises cases where the respondent is unable to fully complete the Diary due to doubts about how to fill in or the meaning of the questions, does not speak Italian or declares that they make most of their expenses every month or less frequently than the survey duration;
- inconvenience of completion methods, which includes notes where the respondent explicitly requests an update of the questionnaire and Diary administration methods, or to be notified earlier about the start of the survey;
- technical aspects, which include cases where respondents make shared expenses with non-cohabiting family members or request a reduction in the effort required to complete the Diary, which is considered too long and tiring.

In the following schema, the distribution of the relevant notes within each subcategory is reported. Given the small number of comments and their uniformity over the two years under analysis, Table 4 has been constructed using the combined data from both years.

Table 4. Macro-categories and related codes identified for the final interview notes

Macro-category	Code	Notes	
		N.	%
Difficulties in contact	Absence of respondents	2	2.8
	Low participation	8	11.3
	Distrust	10	14.1
Difficulties in completion	Clarity of questions	3	4.2
	Language or area issues	7	9.9
	Injury or illness	8	11.3
Inconvenience of completion methods	Requests a digital questionnaire	2	2.8
	Requests advance notice	5	7.0
Technical aspects	Diary too long or complex	19	26.8
	Shared expenses with others	7	9.8
Total		71	100.0

Source: Authors' elaboration on Istat data

The described framework provide valuable insights, useful to ensure that the redesign of the survey methods, currently in the planning phase, represents not just a simple translation of the current Diary onto a digital platform, but an improvement from multiple perspectives. Below are the solutions deemed most effective to address the difficulties encountered by interviewers and respondents (Table 5).

Table 5. Problems encountered by interviewers and respondents and related solutions

Frequent issues	Possible solutions
• Length and complexity of the Diary • Language or area issues • Injury or illness • Clarity of questions • Absence of respondents • Low availability • Distrust and privacy concerns • Request for a digital questionnaire • Request for advance notice • Shared expenses with others	• Preloaded list of items to be entered • Simplification of questions for a cleaner interface • Integration of scanners for reading receipts • Option to choose the language • Inclusion of tooltips and FAQ section • Sending notifications via App • Incentives for completion • Respondent engagement programs • Real-time chat with an operator • Option to choose between the App and the paper Diary • Option to complete the questionnaire via multiple devices (smartphone, tablet and PC) • Allow multiple people from the same household to use the App

Source: Authors' elaboration on Istat data

5. Considerations and future perspectives

The classification described in the previous paragraphs covers a series of problems that the respondents encountered during the interview and the compilation of the Diary, as well as interesting ideas for improving the survey from the perspective of the survey instrument. In fact, from the analysis of the notes, it emerged that the main problems recorded refer, on the one hand, to the structuring of the survey instrument and, on the other, to the time required for its complete compilation since the Diary is considered, by many interviewees, to be too complex and/or too long. Analyzing the point of view of respondents and interviewers, through notes, allows us to identify and better understand the challenges to be addressed in improving data collection, by developing user-friendly systems, and reducing difficulties for respondents. The classification of the problems and suggestions proposed could be used to improve the design of the Diary and make it easier for respondents to fill in while raising interventions aimed at addressing specific difficulties that respondents encounter when recording expenses.

The analysis of the notes and the proposed classification could also serve as a basis for future research on the problems that respondents encounter when compiling the Diary and conducting interviews. Given the potential benefits of gathering a larger number of observations, one possible approach could be to introduce mandatory questions to ensure more comprehensive data collection. This would help to strengthen the robustness and reliability of the findings, improving the validity of the conclusions drawn from the current analysis. By increasing the sample size and ensuring more consistent responses, such an approach could provide a more solid foundation for drawing conclusions and making recommendations. The main objective of this study was to gain an understanding of the perspectives of respondents and interviewers, as well as to analyze their suggestions and proposals. The analysis did not specifically account for demographic variables, as the primary focus was on the qualitative insights derived from the feedback provided. While considering demographic differences could offer valuable perspectives, the scope of this study was limited to the immediate comments collected. Future research could explore these factors further to better understand potential variations in respondents' experiences. To provide a comprehensive overview of the strengths, weaknesses, opportunities, and threats associated with the classification of the notes in the questionnaires and Diary, a SWOT Analysis (Figure 1) was conducted, enabling a deeper assessment of its implications and impact (Pahl and Richter, 2007).

Figure 1. SWOT Analysis on comments provided by interviewers and respondents



Source: Authors' elaboration

Starting from 2023, Istat has been analyzing, based on the study of foreign experiences, the potential offered by technological and methodological innovation, such as the use of mobile devices and dedicated Apps, to optimize the efficiency and accuracy of the survey process. It is important to design and develop an intuitive and user-friendly App that allows households to accurately record their daily expenses, efficiently organizing and managing the data collected to enable in-depth analyses and generate detailed reports.

There is a need to design and invest in more advanced data capture systems, such as web questionnaires and mobile Apps that comply with privacy regulations to ensure the protection of respondents' personal data. From a survey technique perspective, the analysis of the notes confirms the need to reduce the percentage of paper-based data collection forms in favor of data acquisition methods more aligned with the new technologies. It also suggests ideas and small adjustments to ensure the App is accessible to all respondents.

References

D'Amen, B., Pontecorvo, M.E. (2012). L'analisi testuale in rete: un'integrazione degli approcci lessicometrico ed ermeneutico nello studio dei blog Pro-Ana, in *Actes des Jadt 2012. Journées Internationales d'Analyse Statistique des Données Textuelles*, Louvain-la-Neuve: Ciaco, pp. 309-322.

Giuliano, L., La Rocca, G. (2012). *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*. LED Edizioni Universitarie, Milano.

Istat (2016). *Household Budget Survey*. UniData - Bicocca Data Archive, Milan. Study Number SN185. Data file version 1.0.

Istat (2023). *Indagine sulle Spese delle Famiglie.* Available at: https://www.istat.it/it/archivio/71980

Neuendorf, K.A. (2002). *The Content Analysis Guidebook*. Sage Publications, Thousand Oaks, CA.

Pahl, N., Richter, A. (2007). *SWOT Analysis – Idea, Methodology and a Practical Approach*. GRIN Verlag, Norderstedt.

Patton, M.Q. (2002). *Qualitative Research and Evaluation Methods*. Sage Publications, Thousand Oaks, CA.

doi.org/10.26398/asaproc.0098

Combining expert knowledge and data-driven classification to support prediction and validation of census population counts

Angela Chieppa, Antonella Bernardini, Nicoletta Cibella, Fabrizio Solari
ISTAT, Rome, Italy

1. Introduction

Accurate population counts from censuses are necessary for efficient resource distribution, governance, and policy formulation. Therefore, the estimation process must be transparent and easily understandable to non-statisticians to completely fulfil their requirements. With the shift to a permanent census system, population counts in Italy are now produced yearly integrating administrative and survey data in ISTAT statistical registers.

The census process for determining the resident population involves several phases. First, administrative sources are processed within the integrated data base of usual residents (named AIDA) and converted into simplified statistical information called *Signs of Life* (SoL). After selecting only SoL occurrences referring to people eligible to be residents, each individual among the eligible ones is labelled as *resident* or *non-resident* according to a set of classification rules. The rules currently used are based on expert knowledge. Ongoing knowledge discovery from administrative sources and annual census survey data support experts in defining rules. Furthermore, an audit survey is designed for quality assessment. Expert knowledge ensures transparency and comprehension for non-statisticians, while statistical learning could augment this process by discovering additional information not identified by experts.

To take more advantages from the combination of expert and data-driven knowledge, we propose a hybrid classification model. The hybrid model is expected to improve the accuracy of usual population resident estimates and to allow for concrete use of data collected through the audit process, providing effective feedback into the estimation system.

Section 2 describes the methodological context and outlines the theoretical framework. Section 3 is devoted to the proposed architecture for integrating expert knowledge with statistical learning methods in the context of resident population estimation. Finally, Section 4 presents the final remarks and conclusions.

2. Methodological context

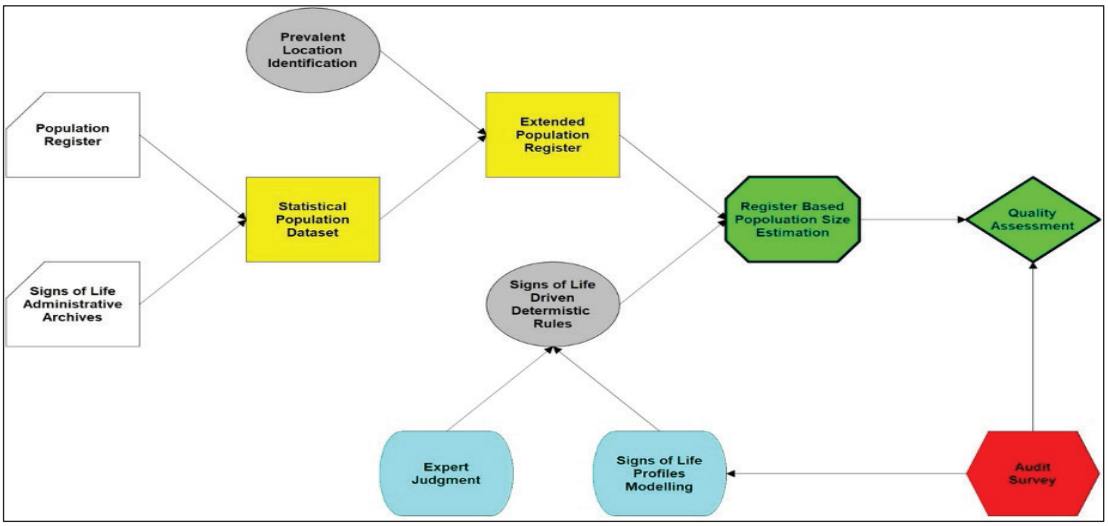
The current strategy of the Italian permanent population census relies on ISTAT statistical registers, which integrate various administrative sources and surveys. A key component is the population register (PR), which consolidates data from municipal registers and other sources. Annual sample surveys enhance the estimation of census results, improving the quality of ISTAT registers in both thematic aspects and coverage. In 2020, ISTAT adopted a fully register-based census approach to estimate the resident population due to the suspension of surveys caused by the COVID-19 emergency. Previously, a combined survey/register approach was used. This new method relies on SoL-based methodology, where SoL, categorized by type, duration, reliability, and geographical relevance, play a crucial role in accurately counting usual residents (UNECE, 2018). The AIDA database plays a pivotal role by centralizing SoL data and cross-referencing PR with other administrative sources.

The estimation process employs deterministic rules that combine results from exploratory data

analysis and statistical modelling with expert knowledge to classify individuals according to residence patterns that emerge from SoL. This classification determines which groups are included in the resident population.

The shift to a fully register-based estimation needed the establishment of a proper statistical framework and outlines the main features of surveys designed to enhance the quality of register-based estimates and provide quality assessment (Solari et al., 2023). Figure 1 provides a visual overview of the framework.

Figure 1. Population size estimation process.



Data resulting from the integration of all the variables in PR and AIDA can be considered as an extended population register (EPR) in which the information related to different SoL coming from the various sources is registered. The EPR is assumed not to suffer from under-coverage, or at most to be affected only by negligible under-coverage compared to over-coverage. This hypothesis was validated with the help of 2021 sample survey data.

The SoL driven deterministic rules classification allows the removal from the EPR of the individuals considered not to belong to the usually resident population, that is adjusting the EPR for over-coverage.The audit survey is specifically designed to assess the quality of register-based population size estimates by measuring estimation errors and collecting data to refine the classification rules, assuming the target population is fully included in the EPR.

Since 2020, experts relied on their thematic knowledge but also on results from a continuous knowledge discovery process on administrative sources to detect individual patterns relevant for estimating the population count. In 2021, ISTAT began using advanced statistical models, such as latent class models, to support experts in defining the classification rules. To further enhance accuracy, ISTAT has been actively testing new techniques and integrating additional data sources into the estimation process.

Recent studies (Laureti et al., 2024) demonstrate the potential of machine learning techniques and novel data sources to improve the accuracy of SoL classification. However, these approaches face significant challenges, particularly with imbalanced datasets. In the Census context, the municipal registers capture a substantially larger population than the subset of individuals with administrative presence but not recorded, leading to classifiers that are biased toward the prevalent group's patterns.

ISTAT is currently implementing and testing the audit survey design. The plan includes a survey on a sample of individuals from the EPR to assess over-coverage, as well as targeted surveys to evaluate the possible population groups underrepresented in administrative data.

3. Hybrid model architecture: Constraints, key challenges and steps

A hybrid model that integrates deterministic rules from expert knowledge with statistical models trained on survey data offers significant advantages. The premise is that the quality of training data and expert knowledge can vary significantly across different population segments. For example, foreigners in specific work sectors are often difficult to be captured by administrative sources, while sample surveys may suffer from non-responses among certain groups, leading to biased training samples. Additionally, expert knowledge can be influenced by personal judgment and experience. The synergy between expert-driven and data-driven approaches helps mitigate these biases: expert judgment is crucial for SoL categories with limited survey data, while advanced statistical techniques are more effective at detecting patterns that might elude expert analysis. Recent literature, driven in part by the proliferation of AI-related methods, highlights the benefits of combining expert knowledge-based approaches with data-driven methods (Wilder et al., 2021; von Rueden et al., 2023).

The hybrid model enhances the census audit framework, described in the previous section, by offering a statistical tool to integrate new survey data, improving the performance of expert classification algorithms. The audit survey is tailored to collect data that enhances the accuracy of critical patterns.The development of a model that integrates expert-driven and data-driven approaches presents several key challenges. First, it is crucial to ensure that both methods are comparable and can be evaluated using the same metrics. Secondly, establishing reliable benchmarks to accurately reflect misclassification rates is essential. Finally, the complexity of integrating these different methodological approaches must be carefully managed to minimize uncertainties.

To address these challenges, it is essential to introduce the statistical setting and then outline the steps required to build the proposed hybrid model. From a statistical point of view, the estimation of the usual resident population is a classification problem. The units to be classified are all the individuals who are eligible to be residents, referring to a specific time t, i.e. all occurrences in the EPR. The target variable is a dichotomous indicator, denoting whether an individual is usually living in Italy. Information about an individual's usual place of residence is a key focus of annual census surveys and is explicitly collected for individuals in the survey samples. Among the current annual surveys, the so-called 'L survey' (where 'L' stands for 'list') collects detailed information on the usual residence of a significant sample of individuals registered in the population register. Additionally, other census surveys may capture information about individuals who are not registered. The steps involved in the hybrid model design are the following:

- the expert-driven approach must be formalized into a rule-based classification model. This involves explicitly defining the variables used and the conditions that govern classification outcomes.
- Statistical learning techniques should be applied to survey data integrated with SoL to develop the most appropriate statistical models. Benchmarks and outcomes must be defined to calculate misclassification metrics, which can be derived from survey results or specific occurrences within administrative data. Careful attention must be given to the definition of training samples, ensuring that the resulting models are robust for all the population profiles and can be interpretable through explicit rules.
- A statistical strategy for integrating expert-driven and data-driven methods must be defined, informed by empirical experimentation and theoretical insights to combine their strengths and mitigate limitations.

The availability of administrative data and census surveys facilitates the creation of data-driven prediction models using data mining techniques. The goal is to predict a dichotomous variable indicating whether an individual is usually living in Italy, leveraging both administrative and survey-derived features.

Key covariates include territorial variables and demographic indicators, validated in recent studies (Bernardini et al., 2024). Among these, the type and pattern of administrative signals (SoL) emerge as the most significant predictors. The quality and availability of data, however, differ substantially between individuals recorded in municipal registers and those unregistered. Municipal registers provide detailed demographic and household variables, such as household composition and foreign citizenship status, and offer a direct measure of the dependent variable, albeit with potential errors. Conversely, for unregistered individuals, fewer covariates are available, and no direct feature representing usual place of living exists, complicating predictive modelling. To address these disparities, the population is segmented into two groups: registered and unregistered individuals, each requiring distinct supervised learning models. For registered individuals, census survey data serves as the primary training set, with outcome variables derived directly from survey results (e.g., detected, not detected, uncertain cases). For unregistered individuals, training samples are defined using a longitudinal approach, analysing administrative signals over multiple years. This segmentation ensures better representation of underrepresented subpopulations and improves model accuracy. The experimental framework is summarized in Table 1.

Table 1- Experimental settings for statistical learning.

Training samples	Main Inputs	Supervising variable
Set 1: individuals recorded in the Population Register at date *t* sampled in Census Survey of same year	Territorial covariates: demographic size of the municipality, degree of urbanization, administrative and territorial reference. Variables from PR: gender, age, citizenship, foreign citizenship status, municipality of residence, number of family members. Variables from AIDA: signal type, signal source, signal location.	Survey outcome (detected, not detected, uncertain outcomes)
Set 2: individuals with consistent SoL at date *t* not recorded in the Population Register at that date	Territorial covariates. Variables from PR *t*: gender, age, citizenship, area/foreign country of birth. Variables from AIDA *t*: signal type, signal source, signal location. Survey Variables: survey outcome (detected, not detected; with the possibility to label uncertain outcomes) of people enumerated in audit or area surveys.	Variables from PR *t*+1: resident flag, municipality of residence. Variables from AIDA *t*+1: signal type, signal source, signal location

Further key insights from first experimentations on available training data highlights the importance of SoL as predictors. Their relevance, however, varies across subpopulations due to interactions with other variables, such as demographic and territorial characteristics. Table 2 illustrates the distribution of SoL profiles among eligible residents and the corresponding Census survey response rates.

Certain subpopulations, such as individuals without SoL or those with weak signals, are at risk of being overshadowed in generalized models due to their low representation. For example, as shown in Table 2, individuals without SoL account for only 1.84% of eligible residents and exhibit low response rates (71.10% for Italians and 35.89% for foreigners), underscoring the need for balanced sampling strategies.

Predictive determinants differ significantly between groups such as foreigners, youth, or individuals with incomplete signals. Tailored models are necessary to capture these variations accurately. For instance, response rates in the census survey vary not only by SoL but also by citizenship, with certain SoL profiles, such as "rental contracts," showing markedly different response rates for Italians (69.17%) versus foreigners (49.63%). This variability necessitates subgroup-specific modelling to avoid misclassification and enhance prediction accuracy.

Table 2 −SoL classes and census survey response rate.

SoL Profiles	Eligible as residents in ISTAT databases		Response rate at L Survey	
	Nr. instances	%	Italians	Foreigners
Steady signs of work/study	31.620.418	52,94%	89,17%	68,32%
Retirement/income source signs	16.926.345	28,34%	90,92%	74,34%
Fiscally dependent family member	4.479.095	7,50%	88,23%	72,34%
Weak signs of work/study	1.932.287	3,24%	85,14%	52,83%
Indirect signs of life (several sources)	1.218.389	2,04%	86,19%	48,35
Rental contract	1.042.950	1,75%	69,17%	49,63%
Signs of university studies	991.543	1,66%	92,17%	74,70%
Signs of work/study episodic	422.545	0,71%	84,49%	55,73%
No signs of life	1.096.850	1,84%	71,10%	35,89%

Source: Data extracted from AIDA database and L census survey, ISTAT 2021

A common challenge in statistics is making a choice among multiple candidate estimation models. Here the term 'model' denotes both statistical models and heuristic classification derived from expert knowledge. Often none of the possible choices can be considered optimal according to some relevant criterion. Model accuracy may be not uniform across the population, each model likely having optimal properties for segments of population representing some critical profiles but not for other critical profiles.

This non-uniform behaviour suggests the need to segment the population and apply tailored models for each group. In this context, hierarchical mixtures of experts (Hastie et al., 2009) offer a statistical solution for handling model selection across different population segments. In the standard mixtures of experts' algorithms, population segmentation is defined using a classification/regression tree, which provides an interpretable structure that can be easily translated into rules for better expert understanding. In similar experimental cases, where a hybrid model combining expert-driven rules and other statistical classifiers is tested, a hierarchical mixture of expert's approach has yielded very promising results (Pradier et al., 2021).

Alternatively, a plausible solution is provided by model averaging (Hastie et al.; 2009), which aims to define an optimal combination of all candidate models. Among model averaging we can distinguish Bayesian and frequentist model averaging. The former is based on posterior probabilities computation and naturally emanating from the Bayesian paradigm, while in the latter, model weights are usually determined to obtain desirable properties of the resulting estimators under repeated sampling and asymptotic optimality. Notice that in the Bayesian context, expert knowledge can be also used to elicit informative priors for the different model parameters.

Within this framework, the expert-driven classification itself can be considered one of the possible expert functions or learning processes. This allows for the integration of expert knowledge as a key component in the overall model, ensuring that both expert insights and data-driven methods contribute to the final classification.

4. Implications and potential benefits of the proposed hybrid approach

This paper presents a theoretical analysis and proposes a hybrid model to enhance the framework for population count estimation. By integrating expert knowledge with statistical learning techniques using survey and administrative data, the model aims to improve the accuracy of usual resident classification. The combination of these approaches is managed through

mechanisms like models averaging or hierarchical mixtures of experts, where combination of classifiers is managed through proper statistical quality functions.

The hybrid approach bridges the gap between deterministic expert rules and statistical learning, allowing audit data to be transformed into actionable insights for population estimation. While offering significant theoretical benefits, the model implementation must balance complexity and interpretability to ensure it remains transparent and understandable to non-statisticians. Additionally, the success of the hybrid model depends on the continuous availability and accuracy of high-quality audit data. Future work will focus on practical implementation and empirical testing to assess its effectiveness and potential for ongoing improvement.

References

Solari, F., Bernardini, A., Cibella, N. (2023). Statistical framework for fully register based population counts, *Metron,* **81**, pp. 109-129.

Bernardini, A., Chieppa, A., Tamburrano, T. (2024). Discovering individual profiles from administrative signs of life useful for the estimation of Census results. *RIEDS - The Italian Journal of Economic, Demographic and Statistical Studies*, SIEDS Societa' Italiana di Economia Demografia e Statistica, **78**(1), pp 15-24.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York, NY.

Laureti Palma, A., Bernardini, A., Cibella, N., De Matteis, G., Solari, F. (2024). Data mining techniques on the administrative data system to enhance the accuracy of the population census counts. *Book of Abstracts of Q2024- European Conference in Quality in Official Statistics 2024*.

McAndrew, T., Wattanachit, N., Gibson, G.C., Reich, N.G. (2020). Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *WIREs Comp Stats. 2021 March-April*. 13(2): e1514.

Pradier, M.F., Zazo, J., Parbhoo, S., Perlis, R.H., Zazzi, M., Doshi-Velez. F. (2021). Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *AMIA Annual Symposium Proceedings*, pp. 525-534.

United Nations Economic Commission for Europe (2018). *Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses*, United Nations, New York and Geneva.

Von Rueden, L. et al. (2023). Informed machine learning – A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1), pp. 614-633.

Wilder, B., Horvitz, E., Kamar, E. (2021). Learning to complement humans. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)*. pp. 1526-1533.

doi.org/10.26398/asaproc.0099

The "group approach" versus the "inverted cascade approach" for calculating the super-dividend test of holding entities

Antares D'Achille, Fiorella Boscaino

National Accounts Directorate, ISTAT, Italy

1. Introduction

ESA 2010 (European System of Accounts[1]) requires that the dividends paid by public companies to the General Government must be super-dividend tested. Following the principle of substance over form, the test measures the portion of the distributed/collected dividends actually attributable in the national accounts as D.42 (dividends deriving from the profitability of the company) instead of F.5 i.e. a return of capital (the super-dividend).

In order to measure the super-dividends, it is necessary first of all to determine the distributable income making a series of adjustments to the financial statements of the companies involved. ESA 2010 indicates a series of adjustments for eliminate all exceptional, atypical accounting items or amounts resulting from the sale/disposal of company assets. In large corporate groups, this process is particularly complex and time consuming because it also requires identifying the possible super-dividends distributed by the sub-holdings in previous fiscal years, making the above cited adjustments to the financial statements of each of years.

The MGDD (Manual on Government Deficit and Debt – Implementation of ESA 2010) provides the option of using the consolidated financial statements in the cases of large corporate groups (group approach), when the use of the individual financial statements (traditional or inverted cascade approach) is burdensome. However, MGDD does not provide specific guidelines or procedures.

Our study moves in this context and provides a comparison between the two approaches: the group approach and the cascade approach, currently used by ISTAT.

The purpose of the research is to identify the pros and cons of both the approaches and to establish if the group approach is as better and/or more efficient method for the assessment of groups' super dividends than the traditional one.

2. Literature review

The most important literature references are ESA 2010 and MGDD. ESA establishes that the "super-dividends are dividends that are large relative to the recent level of dividends and earnings. In order to assess whether the dividends are large, the concept of distributable income is used. (…) The ratio of dividends to distributable income over the recent past is used to assess the plausibility of the current level of dividends. If the level of dividends declared is greatly in excess, the dividends causing the excess are treated as financial transactions and classified as 'super-dividends'. Such super-dividends are treated as the withdrawal of owners' equity from the corporation (F.5)" [2].

Relevant troubles arise when the super-dividends must be assessed for a holding entity, as "the most important component in the profit and loss account of any holding entity is, generally, the dividends it receives from subsidiaries. These dividends are not to be considered as ordinary/entrepreneurial income just because of their mere form. It is necessary that compilers verify whether the dividends received by holding entities are effectively coming from the subsidiaries' distributable income" [3].

[1] Reg. (UE) n. 549/2013.

[2] SEC 2010, par. 4.55 and in the following paragraph, 4.56: "In the case of public corporations, super-dividends are large and irregular payments or payments that exceed the entrepreneurial income of the relevant accounting period, which are funded from accumulated reserves or sales of assets. Super-dividends of public corporations are to be recorded as withdrawal of equity (F.5) for the difference between the payments and the entrepreneurial income of the relevant accounting period (see paragraph 20.206)", p.101.

[3] Task Force on methodological issues Item 5 - Super-dividend test for holdings, share buybacks and recognising a service component in trading holding gains, 2018, point 39, p.6

In this regard, "Eurostat recognises the difficulties linked to the application of the super-dividend test on the results of holding entities (…)." Eurostat recognizes, moreover, that "A plausible solution could be to allow super-dividend testing at the group level (consolidated accounts). This may represent a simplification for compilers".[4]

The issue was introduced in the "Task force on methodological issues" in 2018 when "however, the use of consolidated accounts for the super-dividend test is not explicitly covered by ESA 2010 or the MGDD".[5]

Subsequently, the problematic question of the holdings' super-dividends has been highlighted in the new MGDD revisions. Referring to the issue, the MGDD 2019 states that "this demands an extensive analysis of individual financial statements of the units included in the group which may be time consuming and difficult to quantify. A pragmatic solution for compilers is thus, to use consolidated accounts, which should facilitate the detection of (material) non-ordinary transactions by subsidiaries"[6].

MGDD provides a clear indication for the use of consolidated financial statements, but, once again, it does not provide an operational guide.

3. Methodology

The comparison between consolidated financial statement and parent company financial statement has been studied for a long time, in particular for the assessment of the value relevance of accounting information. In this context, the consolidated financial statement has highlighted more value relevance for investors than individual financial statement[7].

Inverted cascade approach (*source: Individual Financial Statement*)

The Inverted cascade approach[8] requires the analysis of the individual financial statement of the company distributing dividends in order to verify the existence of super-dividends. The super-dividend test needs to identify the configuration of the distributable income[9] by means of a series of adjustments to the financial statement. This procedure, which is not complex in itself, can become particularly burdensome if the company in question is a parent company and the group is particularly complex.

Group Approach (*source: Consolidated Financial Statement*)

The group approach is a method for determining super-dividends provided recently by the MGDD. As highlighted in the Manual, this alternative approach could be adopted if the inverted cascade approach proves too complex to manage and costly to implement. This approach relies on the use of the consolidated financial statements instead of individual financial statements, implicitly targeting groups of companies. Regardless of the complexity of the group and its structure, there is no need to proceed with the retrospective analysis of the subsidiaries because the consolidated financial statements inherently capture the economic and financial results of the all companies.

The Super-dividend

"Super-dividends are dividends that are large relative to the recent level of dividends and earnings." In order to assess whether the dividends are large, ESA refers to the concept of distributable income (DI). The distributable income is calculated as the difference between the net income and a series of adjustments for

[4] Ibidem, points 58, 59, p. 14.
[5] Ibidem, points 64, p. 15.
[6] MGDD 2019 3.5.2.1, point 10, p.160, re-confirmed also in the 2022 version to par. 3.6.2.1, point 11, p.186
[7] Some works, without claiming to be exhaustive: 'Assessing the value relevance of total comprehensive income under IFRS: an empirical evidence from European stock exchanges'; 'The value relevance of other comprehensive income and its components'; 'Value relevance of consolidated versus parent company financial statements: evidence from the largest three European capital markets'; 'The Value Relevance of Comprehensive Income in the European Banking Sector'.
[8] This term has been considered by the analysis of: *Task Force on methodological issues - Item 5 Super-dividend test for holdings, share buybacks and recognising a service component in trading holding gains*.
[9] "Eurostat has consistently considered that the relevant income benchmark is the distributable income. The MGDD only recognises the distributable income as benchmark. Furthermore, this is in line with the provisions of ESA 20.205: The resources available for distribution of dividends is the entrepreneurial income of the corporation.", *ibid,* p.6. Moreover "in practice, the operating profit or surplus (including the net interest charge in order to be consistent with the 'B.4' concept in ESA 2010 as defined in paragraphs 8.26–8.27) of the corporation could be used as a proxy for the entrepreneurial income", p. 186, and "the resources from which dividends have to be paid should neither include the proceeds of sales of assets nor holding gains", p. 185, and footnote nr. 168: "Income statements may strongly differ when established on a national or proper to a corporation basis. In all cases it must refer to the net result of current activity ('continuing' under IFRS), before distribution and income tax, excluding any exceptional transactions generating holding gains or losses", p. 186, MGDD 2022.

eliminating all exceptional, atypical accounting items or amounts resulting from the sale/disposal of company assets.

The super-dividend (SD) is the difference between the dividend paid and the distributable income.

$$\boldsymbol{SD_t^H = Div_t - distr.I_{t-1}}$$

where:

SD = Super-dividend; Div = Dividend ; distr.I = Distributable Income

Determining the $distr.I_{t-1}$ of a company needs of the analysis of its financial statement for the year t-1. Determining the $distr.I_{t-1}$ *of a holding company requires the analysis of if financial statement for the year t-1; the analysis of the financial statement for the year t-2 of all subsidiaries distributing dividends in t-1 and so on, depending of the organization of the group (cascade).* In summary, the most difficult part of this process becomes when we need to analyze a complex group of corporations.

In fact, the complexity lies in identifying the previously included super-dividends along with the adjustments necessary for allocating distributable income. It is necessary to consider all the atypical and extraordinary gains and costs (extraordinary operations) as well as the super-dividend already implicitly included in the dividends received in the past from the subsidiaries.

$$\boldsymbol{Adjustments = SD_{t-1}^I + extraor.op._{t-1}}$$

where:
extraor.op. = Extraordinary operations
Adjustments = Changing on accounting values required by ESA 2010

In order to identify such values, one must work backward in time (through the previous financial periods) as many times as there are existing ownership ties between the reference holding company and the indirectly held subsidiaries that have distributed dividends (to their respective parent companies, i.e., sub-holdings of the group).

$$\boldsymbol{Adjustments = SD_{t-k}^I + extraor.op._{t-1}} \text{ con k = 1, 2, …, n}$$

This approach generates a significant workload for analysis. Therefore, it is relevant to test the group approach.

The super-dividend test is the ratio between dividends paid in t (Div_t) and distributable income of previous accounting year t-1 ($distr.I_{t-1}$).

$$\boldsymbol{SD_t\ test = Div_t / distr.I_{t-1}}$$

4. The results

Our analysis focuses the attention on four greatest public groups that regularly pay dividends to General government and that are the most important "payers" of dividends to General Government[10].

We refer to the following groups:

- Gruppo Poste Italiane;
- CDP (Cassa Depositi e Prestiti);
- ENI;
- ENEL.

The observed fiscal years are 2018-2022[11]. We observe this recent period because of both the availability of the necessary sources (as financial reports and analogue standard analysis based on traditional approach) and the consistency of national accounting information needs.

Considering that the observed period includes the pandemic, the analysis of the results deserves particular attention.

[10] Dividends of the four groups represent about 93,50% of the dividends collected by General Government in the 2023.
[11] The SD_t test is based on the accounting data of the previous year. So SD_{2019} test is the ratio between dividends paid in 2019 and distributable income of 2018 (taking into account that the dividends paid are correlated to the income in the previous year).

Table 1. Super-dividend tests in the two approaches (*per cent*)

		2019	2020	2021	2022	2023
ENI	Cascade	116,98	86,95	626,59	84,87	37,33
	Group	84,85	216,93	-40,55	55,99	26,86
Poste	Cascade	104,16	109,10	72,94	167,65	310,59
	Group	23,40	-43,70	233,75	255,82	12,15
ENEL	Cascade	106,17	90,00	238,20	106,89	175,57
	Group	62,47	78,08	87,04	89,01	139,23
CDP	Cascade	110,38	123,18	103,22	73,01	65,95
	Group	109,66	181,82	120,70	68,98	61,06

The data sources used in the two approaches are different and it cannot be claimed that one provides better results than the other. Both sources and approaches are valid and consistent, even though the resulting outcomes differ. Moreover, it is possible that some impact of Covid-19 on the super-dividend test results but by the observation the results seem erratic in the whole period, mainly depending on the amount of distributable income and dividend policies[12]. There are several elements of differentiation between the two approaches that can explain the different outcomes, which can be summarized as follows:

- *Temporal effect*: Individual financial statements from various previous periods are considered to highlight any super-dividends transferred from subsidiaries to sub-holding and ultimately from the holding to general government. This complex effect is not present in the group approach that uses the consolidated financial statements where the accounting data of all units of the group refer to the same fiscal year.
- *Offsetting effect*: Within the consolidated financial statements, there is offsetting among the various positive and negative economic components of income that affect the capacity and definition of distributable income.

Table 2. Dividend/Net Income ratios in accounting and adjusted values (*absolute values*)

Dividend/Net Income ratio	**2018**	**2019**	**2020**	**2021**	**2022**	**Mean**	**Median**	**ST Dev.**
ENI	0,942	1,034	0,800	0,400	0,563	0,748	0,800	0,26
Poste	0,986	0,911	**1,945**	0,962	0,995	1,160	0,986	0,44
ENEL	0,824	0,696	0,996	0,811	0,568	0,779	0,811	0,16
CDP	**1,104**	0,787	0,800	0,543	0,550	0,757	0,787	0,23
Mean	0,964	0,857	1,135	0,679	0,669	0,861	0,846	0,273
Dividend/Net Income adj. ratio (inverted cascade approach)	**2018**	**2019**	**2020**	**2021**	**2022**	**Mean**	**Median**	**ST Dev.**

[12] Dividend policies could be affected by different issues other than economic performances while the latter taken in consideration is the distributable income. The distributable income is a particular income similar to operating income that excludes the effect of extraordinary/abnormal economic components. Also the groups are bigger than SMEs and so more robust and stable and, maybe, to be owned by GG, could be the impact of pandemic period weaker (the size is relevant for COVID 19 impact just as confirmed by: 'The COVID-19 Shock and Equity Shortfall: Firm-Level Evidence from Italy').

ENI	**1,170**	0,869	**6,274**	0,849	0,384	1,91	0,87	2,5
Poste	**1,042**	**1,191**	**1,690**	**2,170**	**3,106**	1,84	1,69	0,8
ENEL	**1,183**	0,912	**2,725**	**1,104**	**1,839**	1,55	1,18	0,7
CDP	**1,104**	**1,232**	**1,032**	0,730	0,602	0,94	1,03	0,3
Mean	1,125	1,051	2,930	1,213	1,483	1,561	1,194	1,075
Dividend/Net Income adj. ratio (group approach)	**2018**	**2019**	**2020**	**2021**	**2022**	**Mean**	**Median**	**ST Dev.**
ENI	0,846	2,071	**-0,406**	0,558	0,267	0,94	0,70	0,8
Poste	0,234	**-0,477**	**5,416**	**3,311**	0,121	2,27	1,77	2,6
ENEL	0,471	0,488	0,627	0,736	0,970	0,66	0,63	0,2
CDP	0,732	0,951	0,737	0,386	0,486	0,66	0,73	0,2
Mean	0,571	0,758	1,594	1,248	0,461	1,131	0,958	0,947

- Different relevance of the dividend distribution policies between the stages in the inverted cascade approach, which does not take them into account in the consolidated approach (except for the final decision on dividends in favor of the general government)[13].
- The magnitude of income taxes, which could differently affect the values in the two approaches

These differences underscore the need for careful consideration and understanding of the implications of each approach in the analysis of super-dividends.

By the way, we have focused the attention on the implicit coherence between dividends paid and net income, so considering that the dividends are issued when the corporation has obtained a profit and, moreover, the former amount is proportional to the latter.[14]

So we have analyzed the relations between dividends and net income, the latter in the different configurations. Using a ratio between dividends and net income, we could assume that, in general, the ratio must be lower than 1 or close to this value. Over time under normal conditions, dividends paid cannot be higher than profits collected.

The ratio values are considered in accrual basis; the values in bold show those that exceed 1 or have negative sign[15] and therefore are considered less coherent[16]. In particular, the inverted cascade approach shows the greatest number of these values, equal to 14, while the group approach and the accounting data show 4 and 2 cases respectively. Of course, also the mean and median are affected from these values. The mean of means and the mean of medians, between the group approach and the accounting one, are more similar rather than the inverted cascade approach.

In general, the group approach is more similar to the accounting data so it is more coherent with assumption of proportionality between dividends distributed and annual profits.

5. Conclusions

[13] The group is seen as a whole entity so the operations are relevant only if incurred between the group and the others entities out of consolidated perimeter. See: Montrone A., Il bilancio consolidato dei gruppi aziendali, chap. 3.

[14] Actually, dividends are also distributed with the scope of giving a stable return to shareholders, also using distributable reserves.

[15] The negative sign means a net loss. The negative values are not considered in the descriptive statistics because artificially would affect the values (improving mean and median) and, conceptually as net losses, it would be not correct.

[16] The assumption is the entity distributes dividends at most of the income earned. In our assessment we exclude the reserves because in the MGDD, 3.6.2.1., point 7: "large payments exceeding any such special reserve set up for short term adjustment, or evidently made from reserves accumulated from more than one year earlier is considered in substance to be a withdrawal of capital", and so, they are assumed as super-dividends.

The analysis carried out allow us to conclude that the group approach represents a valid and conceptually robust alternative against to the inverted cascade approach for the super-dividend test for complex groups. The group approach is more coherent than the traditional approach because ties the group distributable income to the dividend distribution policy.

In the group approach, the economic performance of each company in the group is relevant. The group approach, surely less complex and detailed but complete, reduces the likelihood of make classification mistakes. In fact, using the consolidated financial statement means having only one source of data that, moreover, is characterized by homogeneous accounting rules and option, updated group perimeter. The group approach also makes it easier to control and to audit both the collection of data and the processing.

Finally, the advantage of using the group approach is evident not only in its practical aspects but also in its greater temporal coherence and immunization from intra-group accounting policies.

The study is limited to a set of Italian holding entities/group, a small number of entities but very relevant for general government accounts in terms of amount distributed.

To strongly assess/support the already clear validity of the group approach it would be useful to extend the observation field to the biggest parent companies that pay dividends both to public and private shareholders.

Moreover, it would be interesting to deepen the analysis of super-dividends for financial institution groups, paying special attention to the effective nature of the transactions recorded in some profit and loss account's items relevant for the calculation of the distributable income.

References

Carletti, E., Oliviero, T., Pagano, M., Pelizzon, L., Subrahmanyam, M. (2020). The COVID-19 shock and equity shortfall: Firm-level evidence from Italy. *The Review of Corporate Finance Studies*, **9** (3), pp. 534-568.

D'Achille, A. (2018). The value relevance of comprehensive income in the European banking sector. *International Review of Business Research Papers*, **14** (1), pp. 94-108.

Devalle, A., Magarini, R. (2012). Assessing the value relevance of total comprehensive income under IFRS: an empirical evidence from European stock exchanges. *International Journal Accounting, Auditing and Performance Evaluation*, **8** (1), pp. 43-68.

Eurostat (2013). *European System of Accounts (ESA 2010)*, Publications Office of the European Union, Luxembourg (LU).

Eurostat (2023). *Manual on Government Deficit and Debt – Implementation of ESA 2010 – 2022 edition*, Publications Office of the European Union, Luxembourg (LU).

Eurostat (2019). *Manual on Government Deficit and Debt – Implementation of ESA 2010 – 2019 edition*, Publications Office of the European Union, Luxembourg (LU).

Eurostat (2018). *Task Force on Methodological Issues, Team Work - Item 5 - Super-Dividend Test for Holdings, Share Buybacks and Recognising a Service Component in Trading Holding Gains*, Luxembourg (LU).

Jahmani, Y., Choi, H.Y., Wu, G.J. (2017). The value relevance of other comprehensive income and its components. *Accounting & Taxation*, **9** (1), pp. 1-11.

Montrone A. (2018). *Il bilancio consolidato dei gruppi aziendali. Metodologia di redazione e di analisi*, FrancoAngeli, Milan (IT).

Muller, V.O. (2011). Value relevance of consolidated versus parent company financial statements: evidence from the largest three European capital markets. *Accounting and Management Information Systems*, **10** (3), pp. 326-350.

doi.org/10.26398/asaproc.00100

Migration as a diffusion process: The case of Bologna municipality

Ida D'Attoma, Camila Cid, GiorgioTassinari
Department of Statistical Sciences, University of Bologna, Bologna, Italy

1. Introduction and theoretical framework

The demographic change is currently taking place, characterized by depopulation in certain areas due to phenomena such as population aging, emigration, and declining fertility rates. Italy is a clear representative of this reality, as measurements and projections from the Italian National Institute of Statistics indicate a progressive decline in population throughout the country (ISTAT, 2023).

In this context, it is interesting to understand why this phenomenon is also present in the northern region of the country, which is highly industrialized, with high levels of quality of life and higher per capita income. According to official predictions, while the northern part of the country does not show as rapid a decline as the central and southern regions of Italy, it is still projected to reduce its population by 2.7 million inhabitants (ISTAT, 2023).

In this regard, we are interested in understanding the depopulation phenomenon, with a special focus on emigration from municipalities with high standards of living, such as the municipality of Bologna. Bologna is the seventh most populous municipality in Italy with about 392,800 residents as of Dec. 31, 2022. Among the features worth mentioning is that Bologna confirms itself as an attractive city, as shown by the migration balance, particularly active with the South and the Islands and with foreign countries. 44 percent of immigrants in Bologna are between the ages of 25 to 34 years old and, among them, more than 80 percent are Italian (Comune di Bologna, 2021).

To study the emigration flow of the inhabitants of Bologna, the diffusionist paradigm was used, which has been widely utilized in marketing but has also been applied in various fields of research (Rogers, 2003).

For this analysis, in a first stage, emigration from Bologna was examined through the Bass diffusion model (Bass, 1969). According to this model, in a market analysis context, the diffusion of new behaviours is driven by two types of communication: external, which includes advertising and public awareness campaigns, and internal, which includes social influence, word-of-mouth, and imitation. These two factors give rise to two categories of adopters: innovators, who are primarily receptive to external sources of information, and imitators, who adopt in response to interactions with earlier adopters (Bass, 1969).

As second step, the analysis was complemented with a more structuralist approach (Palloni, 2001), by estimating regression models to address the association of innovation and imitation coefficients derived from the Bass model with sociodemographic and socioeconomic variables. Regarding the latter, the study focused on the economic conditions of the selected destination, allowing us to elucidate whether factors such as unemployment rate, median income, or property values are determinants in the decision to migrate.

2. Data

For this research, we are working with administrative repeated cross-sectional data from public and official sources. To describe the phenomenon of emigration for the Bologna case, data on emigrants were retrieved from the 'Comune di Bologna' website with reference to the period from 1986 to 2023.

However, for the estimation of the Bass model by destinations, the number of emigrants from the Bologna municipality to different regions of the country between the years 2013 and 2023 was used, because the official sites do not provide sociodemographic and socioeconomic information in the required detail to perform these analyses.

To estimate the regression models, different data sources were employed (see Table 1). In particular, data from Bologna municipality were used to gather information on the sociodemographic characteristics of emigrants from Bologna between the years 2013 and 2023. Additionally, information between the years 2013 and 2023 on the socio-economic variables (i.e., unemployment rate and median annual household income in the destination regions) was obtained from the Istat website. Finally, information regarding real estate prices in the destination regions between the years 2013 and 2023 was obtained from the "Agenzia delle entrate" website.

Table 1. Sources of information

Variable	Description	Source	Available at:
Region	Destination regions of emigrants originating from the municipality of Bologna	Comune di Bologna	https://inumeridibolognametropolitana.it/dati-statistici/emigrati-provincia-di-destinazione-serie-storica
Citizenship*	The percentage of emigrants from Bologna who hold Italian citizenship	Comune di Bologna	https://inumeridibolognametropolitana.it/dati-statistici/emigrati-provincia-di-destinazione-serie-storica
Age*	The percentage of emigrants from Bologna aged between 30 and 39	Comune di Bologna	https://inumeridibolognametropolitana.it/dati-statistici/emigrati-provincia-di-destinazione-serie-storica
Sex*	The percentage of emigrants from Bologna who are male	Comune di Bologna	https://inumeridibolognametropolitana.it/dati-statistici/emigrati-provincia-di-destinazione-serie-storica
Unemployment rate	Average unemployment rates between the years 2013 and 2023 by destination region.	Istat	http://dati.istat.it/index.aspx?queryid=20744
Minimum real estate prices	Average minimum real estate prices between the years 2013 and 2023 (in euros) by destination region.	Agenzia delle entrate	https://www.agenziaentrate.gov.it/portale/web/guest/schede/fabbricatiterreni/omi/banche-dati/quotazioni-immobiliari
Median annual household income	Annual average household income between the years 2013 and 2023 (in euros) by destination region.	Istat	http://dati.istat.it/Index.aspx?QueryId=22919

*Variables in percentages terms were introduced in the model as dummy variables; =1 in correspondence of the percentage of each category of interest in the emigrant population during the analyzed period.

3. Descriptive analysis

According to emigration data from 1986 to 2023, there has been an upward trend, from 8,954 to 12,411 cases. However, this trend has not been constant, showing significant fluctuations, such as in 1996, when there was a decrease with 8,468 cases recorded. Similarly, 9,877 cases were recorded in 2014(Comune di Bologna, 2023).

As for the destinations chosen by emigrants, it emerges that in most cases they opt for other municipalities in Northern Italy. Much less frequently, between 1997 and 2012, Southern Italy stood out, but since 2013 it has been overtaken by destinations abroad (Comune di Bologna, 2023).

Focusing on internal migration, in most cases emigrants choose another city in the same province of Bologna. Second, they move to another region and, to a lesser extent, to another province in the Emilia-Romagna region (see Table 2).

As for the Italian regions, Emilia Romagna stands out with significantly higher values compared to other regions, where the destination province with the highest number of cases throughout the period is Bologna. In second place is Lombardy, which, except for the pandemic years, shows a steady increase over time, reaching its highest value in 2022 with 702 cases (Comune di Bologna, 2023).

Table 2. Residential relocations between Italian regions (From 2013 to 2023)

Years	In the same region				Between different regions		Total	Percentage variation
	Within the same province (Bologna)	%	Between different province	%	Absolute values	%		
2013	5.056	47,9	1.189	11,3	4.316	40,9	10.561	
2014	4.711	47,7	1.137	11,5	4.029	40,8	9.877	-6,9
2015	4.890	48,7	1.116	11,1	4.035	40,2	10.041	1,6
2016	5.083	48,4	1.196	11,4	4.218	40,2	10.497	4,3
2017	5.178	48,9	1.221	11,5	4.196	39,6	10.595	0,9
2018	5.281	49,7	1.219	11,5	4.123	38,8	10.623	0,3
2019	5.882	49,7	1.329	11,2	4.621	39,1	11.832	10,2
2020	5.444	48,3	1.291	11,4	4.542	40,3	11.277	-4,9
2021	5.894	50,0	1.436	12,2	4.456	37,8	11.786	4,3
2022	6.314	50,0	1.496	11,8	4.824	38,2	12.634	6,7
2023	5.961	48,0	1.457	11,7	4.993	40,2	12.411	-1,8

Source: Authors' elaboration based on data from the Municipality of Bologna.

4. Methodology

To study the emigration flow of Bologna inhabitants, the standard Bass model (Bass, 1969) was estimated, which besides being the most well-known diffusion model, is also considered the most parsimonious (Parker, 1994)

The Bass model considers three fundamental parameters for its implementation: firstly, there is the potential market size (m), interpreted as the number of people who could potentially emigrate to a particular region; secondly, there is the coefficient of innovation (p), understood as the probability that a person decides to emigrate to a specific region independently of the influence of others; and finally, the coefficient of imitation (q), interpreted as the probability that a person decides to emigrate to a region due to the influence of others.

The formal representation of the Bass Model is a first-order differential equation as in (1):

$$z'(t) = \left[p + q\frac{z(t)}{m}\right][m - z(t)],\ t > 0 \quad (1)$$

where, in an emigration study context, the variation of emigrants by destination z'(t) over time is proportional to the residual susceptible population of emigrants (m - z(t)) and z(t) represents the cumulative number of emigrants by destination at time t.

The model coefficients were obtained for each of the Italian regions chosen as destination by emigrants from the municipality of Bologna. The software R Studio and the DIMORA package (Zanghi et al. 2023) were used.

Once obtained estimated parameters by destinations, following the approach by Kurdgelashvili et al. (2019), two linear regressions were performed: one where the dependent variable is the

estimated innovation coefficient (p) by different destination regions with socioeconomic variables and potentially sociodemographic variables as independent variables; and another where the dependent variable is the estimated imitation coefficient (q) by different destination regions with socioeconomic variables related to the destination region and potentially sociodemographic variables of the emigrants as independent variables.

5. Preliminary results

The estimated Bass models for destination regions of emigrants from the municipality of Bologna show a high level of goodness of fit across all regions. The values of Relative MSE are significantly less than 1, and the R^2 values are close to 1 (See Table 3).

Results showed that most regions exhibit a higher coefficient of imitation (q) than innovation (p), indicating that the decision to emigrate is more driven by the influence of others who have already emigrated, rather than a personal decision to do so.

Importantly, as reported in Table 3, regions such as Emilia Romagna, Veneto, and abroad show a significantly higher coefficient of imitation (q), suggesting that imitation plays a crucial role in the decision to emigrate in these areas. On the other hand, regions with the highest coefficients of innovation (p) are primarily Calabria, Campania, and Sicily. This indicates that in Southern Italy, there is a greater tendency towards independent emigration decisions not heavily influenced by social factors, which could suggest motivations such as returning to one's region of origin or reuniting with family members.

Table 3. Results of the standard Bass models by destination region

	2013-2023										
REGION	**m**	**S.E. m**	**p-value**	**$\hat{p}$**	**S.E. $\hat{p}$**	**p-value**	**$\hat{q}$**	**S.E. $\hat{q}$**	**p-value**	**R^2**	**RMSE**
Abruzzo	2385	8.73	0.000	0.041	0.000	0.000	0.126	0.001	0.000	0.998	0.005
Basilicata	1498	59.69	0.000	0.029	0.001	0.000	0.039	0.002	0.000	0.999	0.004
Calabria	2655	11.21	0.000	0.065	0.000	0.000	0.065	0.001	0.000	1000	0.001
Campania	4134	3.83	0.000	0.058	0.000	0.000	0.162	0.000	0.000	0.999	0.003
E. Romagna	74781	1.08	0.000	0.048	0.000	0.000	0.338	0.000	0.000	0.994	0.017
Friuli-VG	1583	15.42	0.000	0.032	0.000	0.000	0.113	0.001	0.000	1.000	0.001
Lazio	4002	3.96	0.000	0.054	0.000	0.000	0.171	0.000	0.000	0.999	0.003
Liguria	1848	21.14	0.000	0.037	0.000	0.000	0.074	0.001	0.000	0.999	0.004
Lombardia	7277	2.02	0.000	0.046	0.000	0.000	0.261	0.000	0.000	0.997	0.008
Marche	3006	5.43	0.000	0.047	0.000	0.000	0.147	0.000	0.000	0.998	0.006
Molise	703	60.55	0.000	0.033	0.003	0.000	0.014	0.005	0.000	0.998	0.006
Piemonte	2458	8.24	0.000	0.036	0.000	0.000	0.155	0.001	0.000	0.999	0.004
Puglia	4326	3.19	0.000	0.053	0.000	0.000	0.181	0.000	0.000	0.997	0.008
Sardegna	1835	21.72	0.000	0.038	0.000	0.000	0.073	0.001	0.000	0.999	0.003
Sicilia	3424	4.73	0.000	0.055	0.000	0.000	0.141	0.000	0.000	0.998	0.005
Toscana	3752	4.28	0.000	0.051	0.000	0.000	0.171	0.000	0.000	0.999	0.003
Trent-Alto Adige	1779	28.93	0.000	0.035	0.000	0.000	0.094	0.002	0.000	0.999	0.003
Umbria	1118	27.81	0.000	0.028	0.001	0.000	0.134	0.003	0.000	1.000	0.000
Valle d'Aosta	134	NA	NA	0.022	NA	NA	0.031	NA	NA	0.996	0.013
Veneto	4190	3.46	0.000	0.048	0.000	0.000	0.195	0.000	0.000	0.998	0.005
Estero	14232	1.42	0.000	0.049	0.000	0.000	0.303	0.000	0.000	0.997	0.008

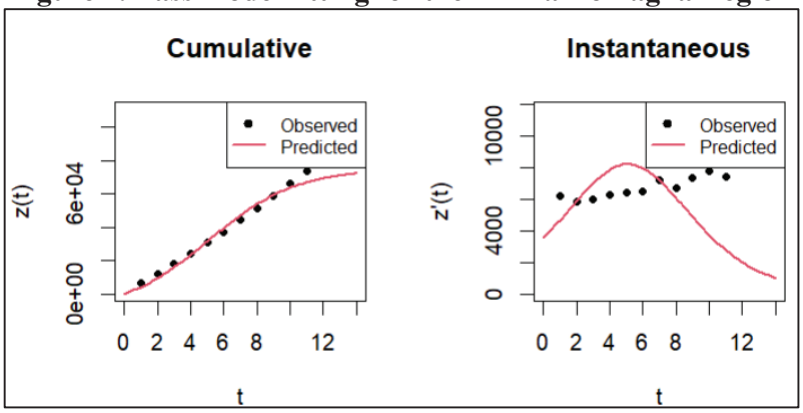
Source: Authors' elaboration based on data from the Municipality of Bologna

With reference to 'Cumulative' and 'Instantaneous' graphs obtained using the standard Bass model to glean insights into the emigration process for each region studied, it is noteworthy that for all analysed regions, the Bass model captures the overall trend of cumulative adoption quite well, fitting the general behaviour of accumulated data over time. This is consistent with the Bass model's definition as a growth model designed to fit a cumulative or integration function of data.

However, the same cannot be said for instantaneous adoption, as in most cases the models fail to fully capture annual fluctuations, resulting in a divergence between observed and predicted values.

Regarding the trajectory of the cumulative curve, most regions exhibit an initial increase in the rate of emigration, followed by stabilization, and in some cases, a slight decrease towards the end of the analysed period. This trajectory can be observed in the Emilia Romagna region (See Figure 1).

Figure 1. Bass model fitting for the Emilia Romagna Region



Source: Authors' elaboration based on data from the Municipality of Bologna

Furthermore, regarding the regression models performed, it is noteworthy that both models showed statistically significant parameters. In the model aiming to explain the innovation coefficient - p- (Model 1), the significant variables are the unemployment rate and the median annual income. While, in the model aiming to explain the imitation coefficient - q- (Model 2), the significant variables are the unemployment rate, minimum real estate prices, and median annual income.

Table 4. Innovation coefficient regression model

Model 1: p coefficient as dependent variable					
Variable	**Estimate**	**S.E.**	**t value**	**Pr(>|t|)**	
Intercept	-1.16e-01	3.20e-02	3.62e+03	2.28e-03	**
Unemployment Rate	4.11e-03	7.00e-04	5.87e+03	2.36e-05	***
Minimum real estate prices	-2.43e-03	6.13e-03	-3.95e-01	6.97e-01	
Median annual household income	3.80e-06	8.50e-04	4.48e+03	3.83e-04	***
Model 2: q coefficient as dependent variable					
Intercept	-9.31e-01	2.75e-01	-3.382	3.80e-03	**
Unemployment Rate	1.70e-02	5.92e-03	2.873	1.10e-02	*
Minimum real estate prices	-1.20e-04	5.47e-05	-2.183	4.43e-02	*
Median annual household income	3.14e-05	7.46e-06	4.202	6.75e-04	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Source: Authors' elaboration of occicial sources of data (see Table 1)

6. Conclusion

In summary, the standard Bass model shows a good fit for all regions of Italy and abroad, as confirmed by the R^2, the Relative MSE, and the accumulation curve. However, the instantaneous adoption rate is not fully captured by the studied models.

Regarding the coefficients obtained through the Bass model, the imitation coefficient is higher in most of the areas studied, with the highest values found in the northern regions compared to the

southern regions of Italy. This indicates that social influence is more decisive in these areas, suggesting different emigration patterns between the two geographical zones.

With reference to the estimated regression models, the socioeconomic variables turned out to be significant for both models. This suggests that emigrants, when deciding the emigration destination seek better living conditions. This is especially true when we consider the variable "Median annual household income", which is significant in both models, and the variable "minimum real estate prices," which is significant for the imitation coefficient. In the case of the unemployment rate, the result is counterintuitive and therefore needs to be studied more thoroughly in future research. This future research could, for example, examine geographical differences in migration decisions.

Disclaimer: This work was funded by the European Union - NextGenerationEU under the National Recovery and Resilience Plan (PNRR) - Mission 4 Education and research - Component 2 From research to business - Investment 1.1 Notice Prin 2022 - DD N. 104 del 2/2/2022, from title [Depopulation Risk: Experiences, Mobility and Subjective Well-being (DREAMS)], proposal code [2022RNKSEL] - CUP [J53D23009480006].

References

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, **15**(5), pp.215–227.

Bitonti, F. (2022). Bass model-based approach to migration. *RIEDS - Rivista Italiana di Economia, Demografia e Statistica*, SIEDS Societa' Italiana di Economia Demografia e Statistica, **76**(2), pp. 4-12.

Carter, A. (2001). Social processes and fertility change: Anthropological perspectives. In J.B. Casterline (ed.), *Diffusion Processes and Fertility Transition: Selected Perspectives,* Washington (DC): National Academies Press (US), pp. 138-178.

Comune di Bologna. (2021). *Piano urbanistico generale: Profilo e conoscenze. Bologna*. Available at https://sit.comune.bologna.it/alfresco/d/d/workspace/SpacesStore/afac50a6-4a08-4eb4-8bbf-cc8a47266b6a/Profilo%20e%20conoscenze_VAR1_ado.pdf (Accessed July 9, 2024)

ISTAT. (2023). *Previsioni della popolazione residente e delle famiglie*. Roma: Istat.

Kurdgelashvili, L., Shih, C.-H., Yang, F., Garg, M. (2019). An empirical analysis of county-level residential PV adoption in California, *Technological Forecasting and Social Change*, Elsevier, **139**(C), pp. 321-333.

Palloni, A. (2001). Diffusion in sociological analysis. In *J.b. Casterline (ed.), Diffusion Processes and Fertility Transition: Selected Perspectives* Washington (DC): National Academies Press (US). pp. 66-114.

Parker, P. (1994). Aggregate diffusion forecasting models in marketing: A critical review. *International Journal of Forecasting*, **10**(2), pp.155-183.

Rogers, E. (2003). *Diffusion of Innovations*. New York: Free Press of Glencoe.

Vijay, M., Muller, E., Bass, F. M. (1995). Diffusion of new products: Empirical generalizations and managerial uses, Marketing Science, INFORMS, **14**(3_supplem), pp. 79-88.

Zanghi, F., Savio, A., Ziliotto, F., Bessi, A. (2023). DIMORA: Diffusion Models R Analysis. *R package version 0.3.6.* Retrieved from https://CRAN.R-project.org/package=DIMORA.

doi.org/10.26398/asaproc.00101

Users beyond "boundaries": Recent experiences of promoting statistical literacy by Istat at the territorial level among non-regular data users

Silvia Da Valle[a], Susi Osti[b], Lorena Carmen Maria Viviano[b]
[a]Istat - Directorate for communication - Management and dissemination of statistical information, Rome, Italy.
[b]Istat - Directorate for communication - Division for coordinating the promotion of statistical literacy and data dissemination at territorial level, Rome, Italy.

1. Introduction: the cultural context

Making data-driven decisions is crucial for everyone, from ordinary citizens to administrators and policymakers. In our present world, however, the amount of data produced and collected has become increasingly vast, leading to what we can refer to as a "data deluge," with the risk of being overwhelmed (The Economist, 2010). In addition, computational tools have become increasingly powerful, and computational capabilities have reached once unattainable levels. Essentially, we have the chance to "*data-ify*" reality representing it through data, using statistical analysis methods, aided by machine learning and artificial intelligence. Some argue that data are even in danger of becoming a kind of 21st-century deity, as they can be interrogated to reveal secrets hidden in our future (Benanti, 2022). Moreover, in this scenario, the scientific method itself may become obsolete. With enough data and the use of correlation search algorithms, data will speak for itself, if humans are removed from interpreting the underlying causes of phenomena and the dynamics of their interactions (Anderson, 2008).

Therefore, knowing how to navigate data and gain deep awareness of statistical methodologies is becoming increasingly crucial. Only in this way will one be able to avoid the risk of delegating decisions to artificially machine-generated predictive models of reality and thus give up one's ability for critical thinking.

To this end, the importance of developing numerical skills to support critical thinking and problem solving has long been recognized at a global level. At the European level in particular, statistical skills are one of the central components in the framework for key competencies for lifelong learning.

Within this cultural framework, the Italian school system has adopted the general guidelines by providing teaching of statistics in curricular programmes, within the mathematical axis[1]. Despite statistics having been a part of the school curriculum for several years, from elementary through to the entire secondary education cycle, the numeracy level of Italians remains inadequate. The results of international student assessment programmes are not comforting for Italians. According to the Programme for International Student Assessment, the OECD-PISA 2022 survey, average math skills in Italy declined between 2018 and 2022, in line with the overall trend among OECD countries. Italy appears to be approaching the OECD average from a position among nations with below-average math skills.

Thus, the signs that can be discerned are not very encouraging overall: Italy is steadily negative within a worsening OECD area (OECD, 2023).

[1] For more in-depth information on the regulations pertaining to the national curriculum directions for kindergarten and the first cycle of education, as well as for the second cycle of education and the school system in general, please consult the Ministry of Education and Merit's information page: https://www.miur.gov.it/web/guest/percorsi-di-studio-e-formazione (Last accessed: 07/24/2024).

This lack of numeracy, which is crucial for active and informed participation in modern society, undoubtedly limits the goal of enhancing the use of data as a tool for understanding reality and supporting informed decision-making processes.

Given its institutional role and mission, the National Institute of Statistics (Istat) in Italy is a key player in the current cultural context. Istat in fact, not only serves as the main producer of official statistics in the country but also plays a crucial role in promoting statistical culture: it supports robust decision-making processes based on quantitative information for all the players of society (citizens, administrators, public and production sectors).

In the context of the most recent strategic directions, Istat is strengthening its role and profiling itself as an advantageous cultural intermediary to provide tools to help interpret phenomena in everyday life, including for non-experts, with a particular focus on younger generations. The challenge for Istat is to support the data value chain (European Commission, 2018) comprehensively, in all its stages: from production to the promotion of the use of official statistics. In response to these new needs, Istat, within the Central Directorate for Communication, has recently established the "Division for coordinating the promotion of statistical literacy and data dissemination at the territorial level", to produce effective actions also at a local level.

The Division's activities include two groups of adult training programs, designed and implemented to expand the audience of data users beyond the "boundaries" of specialist targets, with particular attention on positive and long-lasting effects on younger generations.

This particularly involves activities aimed at promoting statistical culture among teachers at all levels of the school system and targeting managers in the "Third Sector." The actions, despite the difference in recipients, have the same objectives: increasing awareness on a large scale of importance for everyone to have data produced according to quality standards and ensuring that data are used on a daily basis for educational and professional purposes. In this way, data can become an indispensable resource in a changing world and a value added for an increasingly wider audience of users.

2. The experience with teachers in the school

In general, the research community has significant potential to contribute to improving the national school system, facilitating the transfer of knowledge, and actively promoting a widespread scientific culture within schools and among teachers (Istat, 2023). For some time now, Istat has established a dense network of collaborations with educational institutions at all levels, aiming to develop statistical knowledge within the Italian school system, from pre-school to secondary school, in line with the agreement between the Ministry of Education and Merit (MIM), initially signed in 2019 and renewed in 2023 (MIM and Istat, 2023).

The teaching of statistics and probability in Italy faces several challenges, including the lack of an established tradition. Many mathematics teachers lack specific training in statistics and have little direct experience in teaching it. Overcoming this vicious cycle requires that teachers become familiar with statistics and probability, understanding their essence and connection with other mathematical fields, so that they can find pleasure in teaching these areas and can implement them effectively in the classroom (Ottaviani, 2011).

Istat has long recognized schools as an advantageous place to address the issue of statistical literacy. The first schools project, launched in 1987, included a booklet entitled "Statistical Information - let's talk about it with Istat" and a series of slides (Istat, 1986). Subsequently, many other teaching materials were created, all of which were accessible online. The work of Istat's regional offices has been fundamental in developing pathways which are differentiated in terms of content, mode of intervention, and design depth, always seeking the active involvement of teachers.

From September 2021 to June 2024, numerous initiatives were organized to promote data culture for teachers at all school levels, from pre-school to secondary school. During this period, more than 81 events were held, with more than 190 hours of training and the participation of over 4,300 teachers (Da Valle et al., 2024). The aim of these initiatives was twofold: to promote the use

of official statistics in the classroom through data-driven workshop activities and to experiment with new ways to of communicating and disseminating statistics to support teaching.

During events for teachers, not only are the data and tools of official statistics presented, but ideas, activities, and classroom materials are also shared. The proposed activities align with the methodological guidelines of the "4Cs" STEM framework (MIM, 2023), emphasizing the integration of theory and practice and the importance of using data to develop critical thinking skills.

The interventions are custom-designed and implemented in different formats, often in collaboration with cultural institutions, universities, and associations in the area. These partnerships are crucial for identifying training needs, formulating relevant and appropriate proposals for professional development, and amplifying the impact of the interventions.

The initiatives implemented explore various formats: in-service courses, summer schools, and seminars held during dissemination events of national significance. The duration of the interventions is flexible and tailored to the content, with both in-person and online participation options to encourage engagement, even in remote areas.

The content is closely linked to the methodological indications provided by the national guidelines and school curriculum. The working methods include direct experience in line with the STEM approach and cross-disciplinary contamination, even with areas not strictly related to mathematics, such as music, art and literature. Hands-on workshop experiences are offered to familiarize participants with the channels for accessing Istat's official statistical information, in order to promote data-driven teaching to the fullest extent.

All training meetings include moments of dialogue and sharing of classroom experiences, becoming opportunities for exchange to listen to teachers' needs and better tailor Istat's offer of teaching materials and the focus of future meetings.

To strengthen actions at the local level and broaden the impact of its promotional initiatives, Istat has been participating for several years in the Didacta Italy Fair, the leading event on innovation in the Italian education system. This event offers an opportunity to connect not only with teachers but also with school principals, who play a crucial role in planning teacher training.

All these initiatives aim to enhance teachers' statistical skills, contributing to making statistics an integral and dynamic part of the school curriculum.

The effectiveness of Istat's interventions for teachers was evaluated through an impact assessment, after the conclusion of some cycles of meetings considered "barometer initiatives"[2]. The results are highly satisfactory. Specifically, among the respondents (57) who participated in these selected meetings, all rated the potential impact of the training on their work as either 'very much' or 'quite a lot.' Moreover, all participants stated that they would recommend the course to their colleagues.

3. The experience with third sector management

Istat and the National Forum of the Third Sector have embarked on a collaboration aimed at realising a path to promote the culture of statistics for a socially useful territorial action. The activity is part of the initiative "Users. Promotion of Istat's information assets to specific targets in the territory"[3].

The ultimate goal of the initiative is to ensure that the adult target involved can develop their own professional growth from interventions and focuses on the in-depth study of official statistics data, their correct identification and interpretation in order to make conscious choices based on them.

[2] These are, respectively, the last edition (2023-2024 school year) of a consolidated training course, organised since 2015 in collaboration with the 'G. Parenti' Department of Statistics, Informatics and Applications of the University of Florence, and the last two editions (2022-2023 and 2023-2024 school years) of the Istat-Mathesis Summer School, for their importance in terms of innovation.

[3] Structured initiative within the Division for coordinating the promotion of statistical literacy and data dissemination at the territorial level, of the Central Directorate for Communication of the Italian National Institute of Statistics.

The theoretical insights that frame the initiative refer to- the "centrality of the user", as highlighted by Brown (2009) regarding "design thinking", Kolb (1991) and Knowles (2002) with reference to adult learning theories.

The National Forum of the Third Sector is a non-profit organisation and the main unitary representative body of the Italian Third Sector. It consists of thousands of territorial articulations, regional and provincial Forums. It also carries out its mandate in numerous fields: voluntary work, associations, social cooperation, international solidarity, ethical finance, fair trade.

The collaboration established with the National Forum of the Third Sector is addressed to the managers of its numerous member organisations, which are involved in social issues that are reflected in ISTAT surveys. Facilitating knowledge of these surveys and the main ways in which they are disseminated along with including in-depth studies on statistical methodology become opportunities to foster the professional growth of the managers and executives involved in the training.

The promotion path is made up of two parts, which although responding to different objectives, are complementary to each other: e-learning modules and thematic workshops.

The e-learning modules are asynchronous, self-learning on the Forum of the Third Sector MOOC (Massive Open Online Courses) e-learning platform available at https://www.fqts2020.it/ after registration.

Three training modules have been designed and disseminated, each preceded by a useful webinar event to raise awareness among stakeholders:

- Navigating the Istat website in search of data for the Third Sector: through the navigation of the Istat website, the dissemination of data in its main modalities (microdata, databases, press releases, publications) is presented.

- Analysis of official statistics data: basic statistical methodology (data tables and frequencies, graphical representation, main measures of central tendency and variability, representation on maps and construction of infographics).

- Istat, Sistan, Eurostat and contacts in the territory: regulatory framework of the role of Istat and Sistan (Statistics Code, National Statistical Programme). Contact services with Istat through requests for data and information (Contact Centre, Rivaluta), Eurostat databases.

The standard structure of each module foresees a set of materials consisting of a video recording by Istat experts and supporting slides accompanied by bibliographical insights. For each module and for the entire course, there is a self-assessment test with the aim of issuing a certificate.

In terms of duration, the entire course requires a commitment of approximately six hours to watch the videos and an off-line study of approximately fifteen hours.

More than six hundred people are currently enrolled in the modules, which testifies to the interest in the discipline of statistics and official statistics more generally. The three-module course remains available on the platform. Given the large number of learners, it will be interesting to evaluate the level of enjoyment of the course assessed by means of a final questionnaire.

After acquiring common skills, the course continues with several in-depth thematic online workshops, the contents of which focus on the domains characterising the mission of the Forum's member organisations. The workshops enable participants to actively work on the databases, identifying data, commenting on their trends, and carrying out in-depth statistical analyses.

To address potential obstacles to distance training (such as skill gaps, connectivity issues, or distractions), platforms familiar to learners were employed. Additionally, engaging and collaboratively designed support materials were provided to enhance usability, deepen content comprehension, and verify learning outcomes.

To conduct an initial assessment of the training's impact, user satisfaction with the three e-learning modules and the potential application of official statistics in their work were evaluated. This subjective assessment was included in the final questions presented upon completion of each module. To date, 123 participants have responded. On a 5-point Likert scale, approximately 70% indicated that they would make "very much" or "substantial" use of the

new knowledge, while less than 4% reported that they would not use it at all. If confirmed by a larger sample size, these results could be considered more robust.

The collaboration between Istat and the National Third Sector Forum represents an example of useful collaborative work for a specific adult target with the aim of fostering the correct use of official statistics data pursuing a civic and socially useful purpose for the country.

4. Conclusions

Traditional conceptions of data have evolved. Data are no longer just numbers classified into quantitative or categorical types and typically stored in static spreadsheets. Today, data can be dynamic and complex collections, either highly structured or unstructured, such as images and sounds. Data sets are now large and easily accessible. Moreover, the computational power of modern computers makes it possible, even with simple tools accessible to all, to exploit data pools and translate them into knowledge. The opportunities for growth and knowledge are vast and accessible to everyone.

Therefore, it is necessary to promote greater statistical education not only among intensive users (Vichi et al., 2015) but also among less experienced adult users, those "beyond the boundaries". Sound statistical education is essential for social progress, sustainable community development, and the cultural growth of informed citizens who can actively and responsibly contribute to society now and in the future. Only in this way can we aspire to an informed citizenship that is less susceptible to manipulation by sensationalist statistics and less exposed to the risks of superficial use of new artificial intelligence tools.

The experiences presented in this paper were conceived and designed by Istat to address the needs for growth in statistical culture expressed by the identified stakeholders (teachers and management). This approach allowed for the development of an offering that aligns with the target audience's context and real needs while spreading awareness that official statistics can enhance everyone's decision-making abilities in both daily and work life.

In the next three-year period (2024-26), Istat plans to enhance its focus on regional needs by increasing efforts to promote statistical culture (Istat, 2024). This will involve reaching out to new audiences who may be interested in deepening their understanding of official statistics, with a view to personal and professional growth, and creating an indirect spill-over effect to everyone interacting with them.

Potential barriers to participation must be carefully considered when designing new activities. One significant barrier identified in the initiatives described in this paper concerns Istat's visibility as an active institution in promoting statistical culture and adult education. While Istat is widely recognized as the primary producer of official statistics, its role in raising awareness about data culture -an integral part of its third mission- is less well-known, if not entirely overlooked. Another obstacle to meaningful participation lies in the cultural sphere. In Italy, the adult participation rate in educational activities saw an increase in 2022 to 9.6%, yet it remains among the lowest in Europe, where the average is 11.9%. This figure is significantly lower than in Scandinavian countries, such as Sweden (36.2%) and Denmark (27.9%).

Moreover, there is a high variability at the regional level[4]. The territorial dimension is therefore crucial in designing future training interventions aimed at adult targets. In order to expand the audience and extend the initiatives at the territorial level, it will be necessary to adapt to the specific needs of each sector and to collaborate with local stakeholders, working in synergy with the inter-institutional networks that Istat has developed over the years.

References

[4] For details, please refer to https://www.istat.it/wp-content/uploads/2024/04/2.pdf and to https://noi-italia.istat.it/pagina.php?id=3&categoria=5&action=show&L=0

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete *WIRED* https://www.wired.com/2008/06/pb-theory/ (Accessed: 09.07.2024).

Benanti, P. (2022*). Human in the loop - Decisioni umane e intelligenze artificiali*, Mondadori Università.

Brown, T. (2009). *Change by Design: How Design Thinking Can Transform Organizations and Inspire Innovation.* Harper Collins.

Da Valle, S., Bailot, M., Caporrella, N., Fusco, V., Osti, S., Rubbo, C., Spinella, V. (2024). Statistiche ufficiali nelle mani degli insegnanti: esperienze e prospettive di promozione sul territorio. Poster alla Conferenza Nazionale di Statistica 2024. https://www.istat.it/storage/15-Conferenza-nazionale-statistica/poster/07_35_Da_Valle_POSTER.pdf (Accessed: 09.08.2024)

European Commission (2018). *DG CONNECT A European Strategy on the Data Value Chain* https://digital-strategy.ec.europa.eu/en/library/elements-data-value-chain-strategy (Accessed: 24.07.2024).

Istat (1986). Informazione statistica – parliamone con l'Istat Pubblicazione monografica, https://ebiblio.istat.it/digibib/Miscellanea%20Istat%20e%20collaborazioni/AQ10016223Informazione_statistica_parliamone_con_lSTAT.pdf (Accessed: 12.08.2024)

Istat (2023). Il valore sociale della cultura statistica, Collana: *Letture statistiche* – Metodi Pagine: 136 ISBN: 978-88-458-2130-1, https://www.istat.it/produzione-editoriale/il-valore-sociale-della-cultura-statistica/ (Accessed: 09.08.2024)

Jonas, N. (2018). Numeracy practices and numeracy skills among adults. *OECD Education Working Papers (177)*

Knowles, M.S. (2002). *Quando l'adulto impara.* Franco Angeli

Kolb, D. (1981). *Experiential Learning Theory and the Learning Style Inventory*. The Academy of Management Review.

MIM - Ministero dell'Istruzione e del Merito e Istat - Istituto Nazionale di Statistica (2023). Protocollo d'intesa tra MIM e Istat. Per la promozione e la produzione di analisi statistiche nelle istituzioni scolastiche, nella prospettiva dell'inclusione scolastica e sociale, 1.12.2023.

MIM - Ministero dell'Istruzione e del Merito (2023). *Linee guida per le discipline STEM*. Roma, Italia: https://www.miur.gov.it/documents/20182/0/Linee+guida+STEM.pdf/2aa0b11f7609-66ac-3fd8-2c6a03c80f77?version=1.0&t=1698173043586 (Accessed: 29.11.2024)

OECD (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en. (Accessed: 09.07.2024).

Ottaviani, M.G. (2011). Insegnare e apprendere statistica e probabilità a scuola: il problema dell'aggiornamento degli insegnanti. *Periodico di matematiche*, N. 3/2011: 33-44. https://www.mathesisnazionale.it/mathesisbkp/congresso-mathesis/Ottaviani.pdf

The Economist (2010). The data deluge https://www.economist.com/weeklyedition/2010-02-27 (Accessed: 09.07.2024)

Vichi, M., Rosa, V.M.J., Ruane, F. (2015). *The Users of Statistics and their Role in the European Society*. Brussels, Belgium: European Statistical Advisory Committee - ESAC. https://ec.europa.eu/eurostat/documents/735541/6942269/2015_1175_Users+of+statistics+and+their+role+in+European+society.pdf/893231ce-fac3-3951-fe05-5b1c176491c7?t=1634036251674 (Accessed: 09.08.2024)

doi.org/10.26398/asaproc.00102

Conceptualizing and measuring educational poverty through family, school, and environmental opportunities

Cristina Davino[a], Antonio De Falco[b], Rosa Fabbricatore[a], Marco Gherghi[a], Rosaria Romano[a]

[a] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy.
[b] Department of Sociology and Social Research, University of Milano-Bicocca, Milan, Italy.

1. Introduction

Educational opportunities (EO) refer to access and availability of resources that enable individuals to develop skills, knowledge, and abilities necessary for personal and professional development. EO are especially crucial for youth and children, as a deficiency in these opportunities can affect their cognitive and social-emotional development (Heckman et al., 2013). Early unequal EO lead to long-term impacts, including reduced employability, lower salaries, and increased social exclusion (Carneiro & Heckman, 2003).

The literature has identified several factors contributing to EO inequalities among young people, including socio-economic background (Boonk et al., 2018), parental education level (Brunello & Checchi, 2005), quality of schools (Jennings et al., 2015), and living environment conditions (Nieuwenhuis & Hooimeijer, 2016).

Inequalities in EO are closely related to educational poverty (EP), a concept that emerged in the late 1990s within the social sciences. This concept arose from a new understanding of poverty based on its recognition as a multidimensional phenomenon, leading to a new phase of research (Sen & Anand, 1997). This evolution emphasized the importance of considering both economic and educational factors in poverty discourse and shifted the conceptualization of EP to an essential component of human poverty. Consequently, these changes have prompted increased institutional and political attention to EP at both European and national levels, highlighting its significance as a societal issue.

Like other complex phenomena, EP involves multiple dimensions that are not readily observable, making it difficult to define and measure comprehensively. How we define and measure EP affects its recognition, extent, and severity, influencing policy decisions and strategies to reduce vulnerability. Therefore, efforts to define and measure EP must be closely aligned with policy interventions to address issues with significant social impact. From unidimensional approaches focusing on educational qualifications and competencies to multidimensional frameworks considering socio-economic and context-related factors (Battilocchi, 2020), there is still a lack of consensus on how to define and measure EP effectively. A significant step toward a comprehensive definition of EP was made by Save the Children in their report *La lampada di Aladino* (2014). EP is defined as 'the deprivation of children and adolescents from the opportunity to learn, experience, develop, and freely cultivate their capabilities, talents, and aspirations' (p. 4). Drawing on Sen's capabilities theory and Nussbaum's development framework (2011), Save the Children's definition (2015) includes four dimensions of educational deprivation: 1) *learning to understand*, which is related to cognitive and problem-solving skills; 2) *learning to be*, which involves psychological and emotional development; 3) *learning to live together*, which entails the ability to promote social and inter-personal relationships; 4) *learning to lead an autonomous and active life* involving aspects like health, physical integrity, and food security, which are regarded as functional conditions for education and other learning opportunities.

Within this framework, EP is viewed as a deprivation of EO that hinders personal development. This deprivation impacts cultural resources, cognitive abilities, emotional and relational aspects,

and life planning, all of which shape the developmental trajectory of minors (Battilocchi, 2020).

Building on these considerations, our work aims to contribute to the ongoing debate by introducing a novel measure of EP at the individual level. Specifically, we propose a measurement strategy employing a set of indicators to assess the EO available to students across multiple critical domains in which they are embedded. This approach helps define the structure of opportunity constraints concerning educational resources. Further, it allows the evaluation of the impact of different contexts on learning opportunities and defines a multidimensional measure of EP.

The conceptual proposal described in Section 2 is based on a convenience sample (Section 3) of 191 Naples's high school students aged between 16 and 19. This case study serves as a crucial validation of the proposed measure of EP, enabling an optimal configuration of dimensions and indicators useful for measuring EP through a combination of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Due to the lack of space, in this paper (Section 4), we will only present the CFA results that verify the significance of the dimensional structure discovered by EFA and confirm the dimensions and indicators defined in the conceptual section.

2. Conceptual framework

This study is part of the "Measuring and Mapping Poverty Education" project, a research initiative supported by the University Research Funding Program (FRA) of the University of Naples Federico II for 2022−2025. Drawing on the Save the Children framework, which views EP as a lack of opportunities for children and adolescents, we developed a conceptual framework that defines EP as a latent and multidimensional construct measured across three dimensions: family, school, and environment, each with its associated set of items.

Based on the existing literature, we also defined a comprehensive model (Figure 1)[1] with two primary hypotheses to enhance the understanding of EP and its effects on individuals.

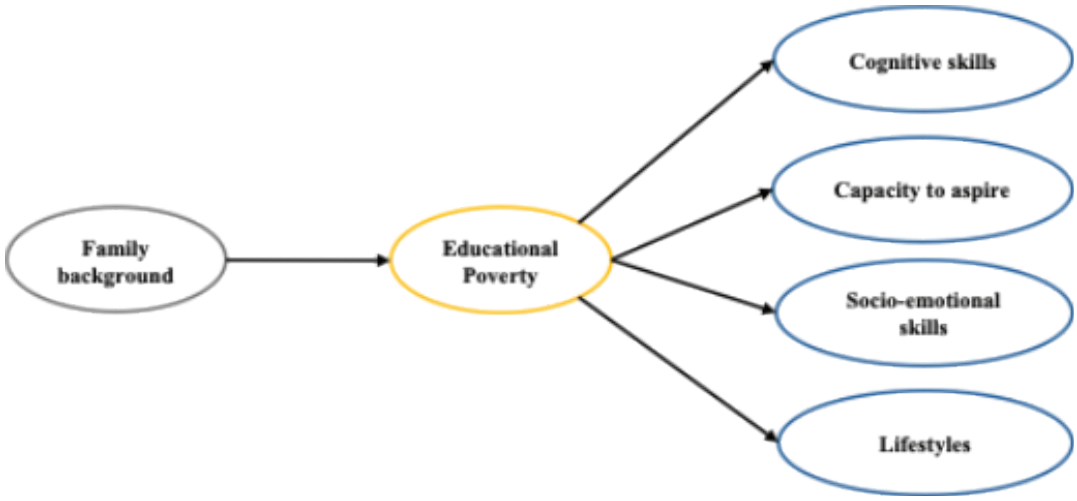


Figure 1 - A conceptual model illustrating the determinants and effects of EP

Firstly, we hypothesized that students' family backgrounds, including parental education levels and occupation, significantly influence the extent of educational deprivation they experience. This influence impacts family resources (Boonk et al., 2018), school choice (Jheng et al., 2022), and neighbourhood of residence (van Ham et al., 2012). Secondly, we posited that EP, characterized by limited access to EO in *family*, *school*, and *environmental contexts*, affects the development of students' cognitive abilities (e.g., school performance) and non-cognitive abilities (e.g., self-esteem,

[1] This model has been presented at the 4th international conference on "Data Science & Social Research" which was held in Naples (Italy) on 25th-27th March 2024. Title of the presentation: Defining a composite measure of Educational Poverty at the individual level: a multidimensional approach (Authors: De Falco, A., Davino, C., Fabbricatore, R., Pratschke, J., Romano R.)

motivation, social skills). By exploring these interconnections, our model aims to elucidate the multidimensional nature of EP and the relationship between the EO available to students and the skills they develop. However, in this work, we will focus solely on measuring EP. Therefore, providing a detailed description of the scheme falls beyond the scope of this paper.

Table 1 - Indicators of EP grouped by the Family, School, and Environment domains

Family	School	Environment	Additional variables
1.1 [BOOKS] Do you have the necessary books for your school activities? 1.2 [MAT_SCOL] Do you have all the school supplies (notebooks, drawing paper, calculators, dictionaries, etc.) you need to do homework and classroom activities? 1.3 [TIME_STUD] Do you usually manage to dedicate the necessary time to study when you are at home?	2.1 [DIGIT] Do you use digital devices (tablet, PC, etc.) for educational activities at school? 2.2 [LIBRARY] Are there libraries at school that you can use?	3.1 [SOCIAL] Are there enough places or spaces in your neighbourhood where kids your age can socialize? 3.2 [GREEN_AREAS] In the area where you live, are there parks, gardens, or other public green spaces accessible within a 15-minute walk?	4.1 [GEND] Gender 4.2 [SCHOOL_ATT] Type of school attended 4.3 [IT_GRAD] Grade in Italian 4.4 [MAT_GRAD] Grade in Mathematics 4.5 Having incurred school debts [SCH_DEBTS] 4.6 [HELD_BACK] Having been held back a grade
When you are at home: 1.4 [SPACE_STUD] Do you have a quiet space to study? 1.5 [DESK_STUD] Do you have your own desk to study? 1.6 [PC_TAB] Do you have access to a computer and/or tablet for studying? 1.7 [SUPP_PEOPLE] Do you receive support and/or help with homework? 1.8 [BOOK_HOME] Excluding textbooks, how many books are there in the house?	Does your school offer you the opportunity to: 2.3 [CORS] Attend recovery or reinforcement courses? 2.4 [LAB] Participate in extra-curricular laboratory activities (e.g., programming courses, foreign languages, etc.)? 2.5 [CULT_TRIPS] Take part in cultural trips or organized outings? 2.6 [PRESCHOOL] Did you attend kindergarten (preschool)?	Can you easily reach from the area where you live 3.3 [LIBR_ACC] Libraries 3.4 [THEAT_ACC] Theatres 3.5 [CINE_ACC] Cinemas 3.6 [MUS_ACC] Museums 3.7 [SPORTC_ACC] Sports centers (gyms, pools, etc.)	4.7 [OCCUP_FAT] Father's occupation 4.8 [EDUCAT_FAT] Father's educational attainment 4.9 [OCCUP_MOT] Mother's occupation 4.10 [EDUCAT_MOT] Mother's educational attainment
During the year, do you have the opportunity to: 1.9 [TRAV_FAM] Travel with family 1.10 [TRAV_AMI] Travel with friends 1.11 [MUSEUMS_EXHIB] Visit museums, exhibitions 1.12 [ARCH_SITES] Visit archaeological sites, monuments 1.13 [CINE] Go to the cinema 1.14 [THEATER] Go to the theatre 1.15 [CONCERT] Attend music concerts 1.16 [HOBBY] Pursue hobbies in your free time (e.g., photography) 1.17 [SPORT] Engage in sports			

Regarding the sets of variables used to define the EP construct, the family dimension includes indicators that evaluate access to books and educational materials, time available for study, dedicated study space (equipped with a desk, internet, and a device), and support for academic tasks. Additional indicators assess students' ability to participate in extra-curricular and leisure activities, such as traveling, visiting cultural sites (e.g., museums, exhibitions, theatres, cinemas, concerts), and engaging in hobbies and sports. In the school dimension, indicators assess the availability of educational resources within the school. This includes the use of digital tools (e.g., tablets and computers) for learning, access to the library for study, availability of remedial and advanced courses, and participation in after-school workshops (e.g., language or programming classes). Moreover, indicators evaluate involvement in educational trips and museum visits and whether students attended kindergarten, which provides early opportunities for social, relational, and physical development. Lastly, the environment dimension includes indicators that measure the availability of resources such as public green spaces, libraries, theatres, cinemas, museums, and sports facilities accessible to students in their area of residence.

For the identification of these indicators, we relied on two primary sources: the Save the Children report (2014) and the 2023 ISTAT survey 'Bambini e ragazzi: comportamenti, atteggiamenti e progetti futuri'[2]. Additional indicators were developed ad hoc to capture specific

[2] https://www.istat.it/informazioni-sulla-rilevazione/indagine-su-bambini-e-ragazzi-comportamenti-atteggiamenti-e-progetti-futuri/

aspects of EP not addressed by these sources. All the developed indicators refer to educational resources that, as highlighted in the literature, play a crucial role in the cognitive and social-emotional development of young people. Table 1 details the indicators/questions for each proposed dimension of EP (in brackets labels used in graphs and tables) and additional respondent characteristics.

3. Data and methods

Starting from the conceptual framework outlined in the previous section, we developed a questionnaire to investigate EP individually.

From January to March 2024, we surveyed a convenience sample of high school students in Naples, encompassing those in their third to fifth years. The students were recruited from schools that participated in open day events at one of the University of Naples Federico II departments, as well as during laboratory activities conducted in several Neapolitan schools by the researchers involved in the study. Following a rigorous selection process, 100 students who did not pass an attention check were excluded from further analysis, resulting in a final sample of 191 students: 111 identified as female, 77 as male, and 3 as non-binary. This meticulous approach ensures the accuracy and representativeness of our analysis, providing a robust foundation for our findings. The questionnaire has been structured into sections dedicated to collecting information on the three dimensions of EO (family, school, environment) and individual characteristics such as family background and educational outcomes.

The analysis strategy unfolded in multiple phases. Initially, a univariate statistical examination was conducted on all binary indicators to identify those with sufficient variability. Exploratory Factor Analysis (EFA) was used to assess the tetrachoric correlation matrix, allowing for selecting a cohesive set of indicators for computing scores across the three EO dimensions. Subsequently, Confirmatory Factor Analysis (CFA) was then employed to validate the dimensional structure identified by EFA.

4. Results

Factor loadings obtained through the CFA are shown in Table 2, and they confirm the three-dimensional structure proposed to operationalize EP.

Table 2 – Factor loadings estimate from CFA

Dimension	Indicator	Estimate	SE	t-values	p-values	ci.lower	ci.upper
Family							
	BOOKS	0.530	0.104	5.106	0.000	0.327	0.733
	MAT_SCOL	0.622	0.126	4.936	0.000	0.375	0.870
	PC_TAB	0.658	0.125	5.251	0.000	0.413	0.904
	BOOK_HOME	0.507	0.093	5.475	0.000	0.326	0.689
	TRAV_FAM	0.490	0.099	4.955	0.000	0.296	0.684
	TRAV_AMI	0.587	0.099	5.934	0.000	0.393	0.781
	MUSEUMS_EXHIB	0.625	0.103	6.097	0.000	0.424	0.826
	ARCH_SITES	0.649	0.089	7.287	0.000	0.475	0.824
	THEATER	0.811	0.077	10.591	0.000	0.661	0.961
	CONCERT	0.748	0.083	9.030	0.000	0.585	0.910
	HOBBY	0.617	0.101	6.123	0.000	0.419	0.814
	SPORT	0.503	0.100	5.016	0.000	0.306	0.699
School							
	LIBRARY	0.610	0.127	4.819	0.000	0.362	0.858
	CORS	0.591	0.131	4.496	0.000	0.333	0.848
	LAB	0.942	0.181	5.213	0.000	0.588	1.296

Environment							
	SOCIAL	0.502	0.087	5.764	0.000	0.331	0.672
	GREEN_AREAS	0.450	0.105	4.283	0.000	0.244	0.656
	LIBR_ACC	0.545	0.082	6.617	0.000	0.383	0.706
	THEAT_ACC	0.914	0.039	23.426	0.000	0.837	0.990
	CINE_ACC	0.905	0.045	20.195	0.000	0.817	0.992
	MUS_ACC	0.872	0.046	18.851	0.000	0.781	0.962
	SPORTC_ACC	0.508	0.117	4.345	0.000	0.279	0.738

Model parameters for the paths of domains to items are all significant ("P(>|t|)" = 0). The fit of the model can be considered satisfactory: the Comparative Fit Index is 0.95; the Tucker- Lewis Index is 0.95; and the Root Mean Square Error of Approximation is 0.04, with a confidence interval of 0.03 to 0.05. Further developments of the analysis could consider the computation of a global EP index through the scores identified for each of the three dimensions. Moreover, it is also interesting to identify a possible heterogeneity in the EP distribution by grouping individuals according to similar patterns across the three dimensions.

5. Conclusions

This study proposed a method to measure educational poverty among adolescents aged 15-19, focusing on the accessibility and utilization of educational opportunities. It identified three main domains of opportunities: those provided by the family context, the school, and the social environment. A conceptual model was introduced, suggesting that educational poverty, defined in terms of opportunities, is influenced by the family context in which the adolescent lives and grows. It impacts both cognitive and non-cognitive skills.

The study's results concentrated on the construct of opportunities, measured by 30 indicators. Of these, only 22 demonstrated strong discriminatory and measurement capabilities across the domains considered. Factorial analyses confirmed a three-dimensional structure, underscoring the multidimensional nature of educational poverty and the complexity of the issue. However, some potential limitations of the factor analysis should be acknowledged, particularly regarding the impact of the sample size and its convenience nature on the robustness and generalizability of the results. A larger and more representative sample in future studies could enhance the stability of the factor structure and the external validity of the findings.

Future research will also focus on estimating the entire conceptual model to better understand the network of relationships between educational poverty and other related constructs. This information will be crucial for stakeholders and authorities to devise interventions to reduce educational poverty disparities.

Acknowledgment

Rosa Fabbricatore acknowledges the financial support provided by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP E63C22002140007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

This work was also supported by the University Research Funding Program (FRA) 2022 of the University of Naples Federico II, with the contribution of the Compagnia San Paolo (Title of the project: Measuring and Mapping Educational Poverty"; principal investigator: Cristina Davino).

References

Battilocchi, G.L. (2020). Educational poverty in Italy: Concepts, measures and policies. *Central European Journal of Educational Research*, **2**(1), pp. 1-10.

Boonk, L., Gijselaers, H.J., Ritzen, H., Brand-Gruwel, S. (2018). A review of the relationship between parental involvement indicators and academic achievement. *Educational Research Review*, **24**, pp. 10-30.

Brunello, G., Checchi, D. (2005). School quality and family background in Italy. *Economics of Education Review*, **24**(5), pp. 563-577.

Carneiro, P., Heckman, J. (2003). Human capital policy, in *Inequality in America: What Role for Human Capital Policies?* eds. J. Heckman A. Krueger, MIT Press, Cambridge, (MA), pp. 77-239.

Heckman, J., Pinto, R., Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, **103**(6), pp. 2052-2086.

Jennings, J.L., Deming, D., Jencks, C., Lopuch, M., Schueler, B.E. (2015). Do differences in school quality matter more than we thought? New evidence on educational opportunity in the twenty-first century. *Sociology of Education*, **88**(1), pp. 56-82.

Jheng, Y.J., Lin, C.W., Chang, J.C.C., Liao, Y.K. (2022). Who is able to choose? A meta-analysis and systematic review of the effects of family socio-economic status on school choice. *International Journal of Educational Research*, **112**, 101943.

Nieuwenhuis, J., Hooimeijer, P. (2016). The association between neighbourhoods and educational achievement, a systematic review and meta-analysis. *Journal of Housing and the Built Environment*, **31**, pp. 321-347.

Pratesi, M. (2022). Povertà educativa: perchè e come misurarla anche a livello territoriale. Alcune analisi e proposte. *Politiche Sociali*, **3**, pp. 373-394.

Save the Children (2014). *La Lampada di Aladino. L'indice di Save the Children per Misurare Le Povertà Educative e Illuminare Il Futuro dei Bambini in Italia*. Save the Children Italia.

Sen, A., Anand, S. (1997). Concepts of human development and poverty: A multidimensional perspective, in *Poverty and Human Development: Human Development Papers,* ed. United Nations Development Programme, pp. 1-20.

van Ham, M., Manley, D., Bailey, N., Simpson, L., Maclennan, D. (2012). *Understanding Neighbourhood Dynamics: New Insights for Neighbourhood Effects Research*. Springer, Dordrecht, (NL).

doi.org/10.26398/asaproc.00103

The Italian NEETs after the COVID-19: What should we expect for the future? Different approaches to forecasts

Giovanni De Luca, Antonella Rocca, Paolo Mazzocchi, Claudio Quintano
Department of Management and Quantitative Studies, University of Naples Parthenope, Naples, Italy.

1. Introduction

The issue of young people being Not in Employment, Education, and Training (NEET) is very challenging, at least for countries like Italy, which from 2014 to 2023 showed the highest rates across European countries. The causes of this outcome are known and mainly linked to the labour market conditions and education system characteristics (Pastore et al., 2021, 2022). Regarding the first aspect, Italy has high unemployment levels and a high intergenerational gap when comparing young people's condition with their adult peers. From an institutional perspective, the organizations that should help young people find a job are underdeveloped and inefficient in the sense that they cannot provide adequate ad hoc services to help young people when they enter the labour market. Concerning education, the share of early school leavers is higher than the EU average (in 2022, in Italy, it was 11.5% against an EU-27 average of 9.7% for the age class 15-24) while the share of young people with a tertiary degree is still one of the lowest one, at the EU level (in the same year, 29.2% in Italy, against an EU average of 42.0%; only Romania performed worse, with 24.7%).

The NEET phenomenon has been investigated mainly at the individual level, with the scope to understand the personal and familiar factors connected to falling into this trap, while at the macro-economic level, the attention has been mainly focused on unemployment (Alfieri et al., 2015; Quintano et al., 2018).

The NEET indicator, involving in its definition inactive and unemployed, is more able to capture the attitude of young people to remain unproductive and excluded from the labour market. Consequently, it is the only indicator directly referred to young people included in the UN Agenda of the Sustainable Development Goals (Goal 8 on Decent Work and Economic Growth). The economic literature has widely demonstrated how the NEET phenomenon results from different causes requiring specific interventions. Besides the difficulties in finding a job, being a NEET may depend on the need for care for elderly people or children or health problems. In this article we want to test if, at least among most young people, being a NEET is mainly a problem linked to the scarce availability of jobs or other labour market structural problems. In the affirmative case, the determinants of the NEET status should be the same among inactive and unemployed young people. Studies analyzing the NEET phenomenon from a longitudinal perspective are still limited in the economic literature and refer to very short periods due to the lack of adequate data (De Luca et al., 2019, 2020). The NEET indicator is indeed a recent one. It was introduced in the UK in the late 1990s and referred only to young people aged 16-18 to indicate their disengagement (Mascherini, 2019). In the subsequent years, the age class of reference was prolonged first to 24, after to 29, and finally to 34 years, given the increase in the number of years people employ on average in education and the long time spent in school to work transition, until the attainment of the first regular job (Pastore et al., 2021, 2022).

Consequently, the first attempts to monitor this phenomenon refer to the 2000s. To analyze from a longitudinal perspective the NEET rates, using the time series methodologies, De Luca et al. (2019, 2020) proceeded to construct an ad hoc time series from the Labor Force Survey (LFS), the Eurostat survey finalized to collect information on individuals' professional condition and characteristics. The authors constructed four-week indicators to monitor labour market conditions

using the specific interview week information.

This paper aims to use quarterly indicators on the NEET rates from 2004 to 2022. This time interval allows fewer observations (76), but it shows several advantages.
First, quarterly data is more robust than those based on 4-week time intervals. Second, a similar period allows us to analyze the association between the NEET rates and several other relevant indicators publicly available from Eurostat with the same periodicity. The reference is to GDP, labour productivity, and indicators related to the level of education attained. The age class chosen in this analysis is 15-34 years.

Therefore, using quarterly data, we aim to analyze the association between the indicators of disengagement from the labour market, that is, the NEET rates, youth unemployment, and inactivity, and some essential macroeconomic indicators connected to the national economic framework with the scope to identify the main determinants for each of these indicators of youth exclusion from the labour market (NEET, inactivity and unemployment) and to test the forecast capacity of two different analytical approaches, that is the traditional regression models and the random forest machine learning algorithm.

Regarding the first objective, the analysis will involve macroeconomic indicators, such as GDP at market price and labour productivity, and education system indicators, such as the share of early school leavers and tertiary education. If different paths and determinants are associated to NEETs, inactivity and unemployment rates, we can identify the peculiarities associated with each of them and propose specific interventions to act on the NEET rates overall, but even on their specific components (inactivity and unemployment rates). Finally, with reference to the second objective, the comparison of forecast capacity between two different methodological approaches, we aim to identify the best predictivity model. Few studies, until now, have compared these two different methodological approaches in the analysis of the NEET phenomenon (see for example, Berigel et al., 2024; Rahmani et al, 2024). The novelty of this research is to add to the study of the NEETs, the analysis of unemployment and inactivity, to investigate if different patterns emerge.

2. Methodology

The outcome variables in this analysis are the following indicators:

NEET = (n. unemployed + n. inactive not in education) / total population

INACTIVITY = n. inactive not in education / total population

UNEMPLOYMENT = n. unemployed / (unemployed + employed)

The age class considered for all the indicators is 15-34 years. The time series for these indicators are obtained through *ad hoc* elaborations from the LFS microdata, using the sample weights to infer these results to the population correctly.

For each of these indicators, starting from the independent variables identified in Table 1, through a stepwise model regression for time series, we identified the best set of predictors, and then we moved to:
- Identify for each dependent variable, the most important predictors and compare the three different models' specifications
- Compare the goodness of fit and predictability for the three different outcome variables

We employ two methodological approaches. The first one is the consolidated regression approach for time series. In the second approach, we use the random forest machine learning algorithm, which consists of the advancement of the regression tree, based on random sampling and bagging techniques and on the principle of obtaining accurate classifications by reducing the variance. The term "forest" indicates that there are several trees. The "random" term shows that each tree is only trained on a random subset of samples drawn from the training set (with repetition)

and, possibly, a random subset of features. In other words, the "random" part is needed because otherwise, the trees would be so similar that there would be no advantage in having a "forest". Random forest algorithm builds a class prediction using the majority votes of the decision trees created in a classification problem. Combining majority-voted class estimates of the individually created decision trees allows for identifying the final class.

As the Dickey-Fuller test for stationarity led us to accept the null hypothesis of not stationarity, all the indicators were considered in terms of first differences.

The model fit and predictability capacities were assessed through the RMSE and MAE indicators:

$$\text{Root-Mean-Square Error (RMSE)} = \sqrt{\frac{1}{n}\sum e_t^2} \quad (1)$$

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n}\sum |e_t| \quad (2)$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \frac{1}{n}\sum \frac{|e_t|}{a_t} \quad (3)$$

where e_t are the error terms, that is the difference between the actual and the fitted or forecasted values.

Forecasts are based on the last four years (16 observations), corresponding to the period from 2019 to 2022, a very challenging period, characterized by the socio-economic shock of the COVID-19 pandemic.

3. Results

Table 1 shows the variables used in our analysis. Figure 1 displays the path of the three outcome variables analyzed over time, with the application of a 3-term centered moving average to smooth the irregularities of the indicators.

Table 1 – Variables considered in the study.

Variables	Definition	Identification based on LFS Variables or external source
Dependent variables		
NEET rate	age 15-34 – elaborations from LFS	ilostat>1 & educstat=2 & couratt=2 on total young people
Inactivity rate	Age 15-34 – elaborations from LFS	ilostat=3 & educstat=2 & couratt=2 on total young people
Unemployment rate	Age 15-24 – elaborations from LFS	ilostat=2 divided by the labour forces (ilostat<3)
Independent variables		
Real labor productivity per person employed	% change compared to the same period in the previous year	Eurostat online database
Gross domestic product at market prices	Chain linked volumes, index 2010=100, unadjusted data	Eurostat online database
Temporary employees as a percentage of the total number of employees by sex and age (%)	15-29 years	Eurostat online database
Share of early school leavers	age 15-34 elab from LFS	hatlev1d=="L", ISCED 0-2
Share of tertiary educated	age 25-24 elab from LFS	hatlev1d=="H", ISCED 5-8
Youth Guarantee(*)	0 until Q1 2014, 1 from Q2 2014	-
Citizenship income(*)	0 until Q1 2019 and from Q2 2024, 1 in the middle	-

(*) For these two dummy variables, the transformation in terms of first differences was not applied.

Table 2 reports the determinants of being a NEET, inactive, and unemployed based on the regression model. The covariates are selected for each model through a stepwise algorithm. This latter is an algorithm that, among a certain number of possible regressors, at each step adds or drops one of them with the scope to identify the best model specification in terms of maximization of the

F statistic. Figure 2 shows the key predictors as identified by the random forest algorithm. Figure 3 compares the in-sample and out-sample estimates with the empirical series to assess the goodness of fit and the models' prediction capacity. Finally, Table 3 shows the RMSE and MAE values to compare the goodness of fit of the models identified further.

Figure 1 shows that the unemployment rate has a more cyclical path in comparison to the inactivity rate, while the NEET is in the middle. However, even if less pronounced, the peaks for the inactivity rates reflect at least in part those of the unemployment rates.

The main evidence from Table 2 is:

- The changes in NEET and INACTIVITY rates appear to be associated with the changes in *Temporary employment* at both lags 1 and 4, and *Early school leavers* and *GDP* at lag 4. Moreover, INACTIVITY rate changes are also affected by the variables *Tertiary educated*$_{t-1}$ and *Early school leavers*$_{t-1}$.
- The one-quarter lagged variables influencing the changes in the UNEMPLOYMENT rate are *Labour productivity*, *Tertiary educated*, and *Early School Leavers*. Finally, the variable *Temporary employment*$_{t-4}$ also affects the dynamics of unemployment.

Figure 1 – Moving averages of 3 elements for the NEET, inactivity, and unemployment rates. Age class 15-34.

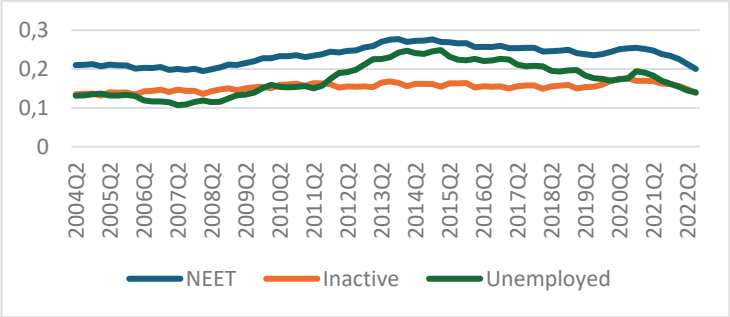
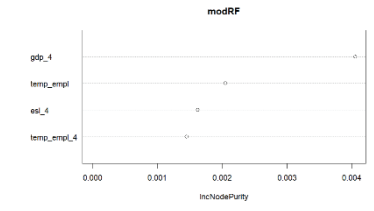
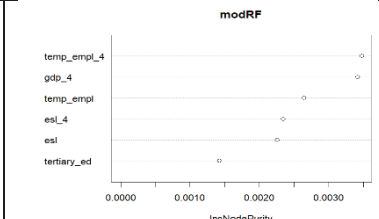
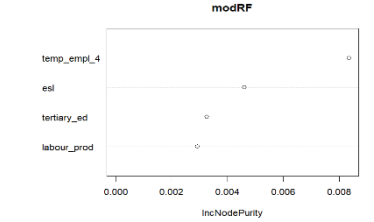


Table 2- Determinants of NEET, inactivity, and unemployment conditions.

	NEET$_t$	INACTIVITY$_t$	UNEMPLOYMENT$_t$
Labour productivity$_{t-1}$			-0.0015***
Temporary employment $_{t-1}$	0.0025***	0.0025***	
Tertiary educated $_{t-1}$		-0.5471**	0.8287***
Early school leavers $_{t-1}$		-0.6966**	0.6733**
Temporary employment $_{t-4}$	0.0022***	0.0036***	-0.0046***
Early school leavers $_{t-4}$	0.4416*	0.5669*	
GDP $_{t-4}$	-0.0015***	-0.0014***	
Intercept	-0.0010	-0.0013	0.0012
Adj R^2	0.4839	0.5225	0.4344
F	18.35***	14.49***	15.21***
N	75	75	75

***' 0.01 '**' 0.05 '*' 0.1.

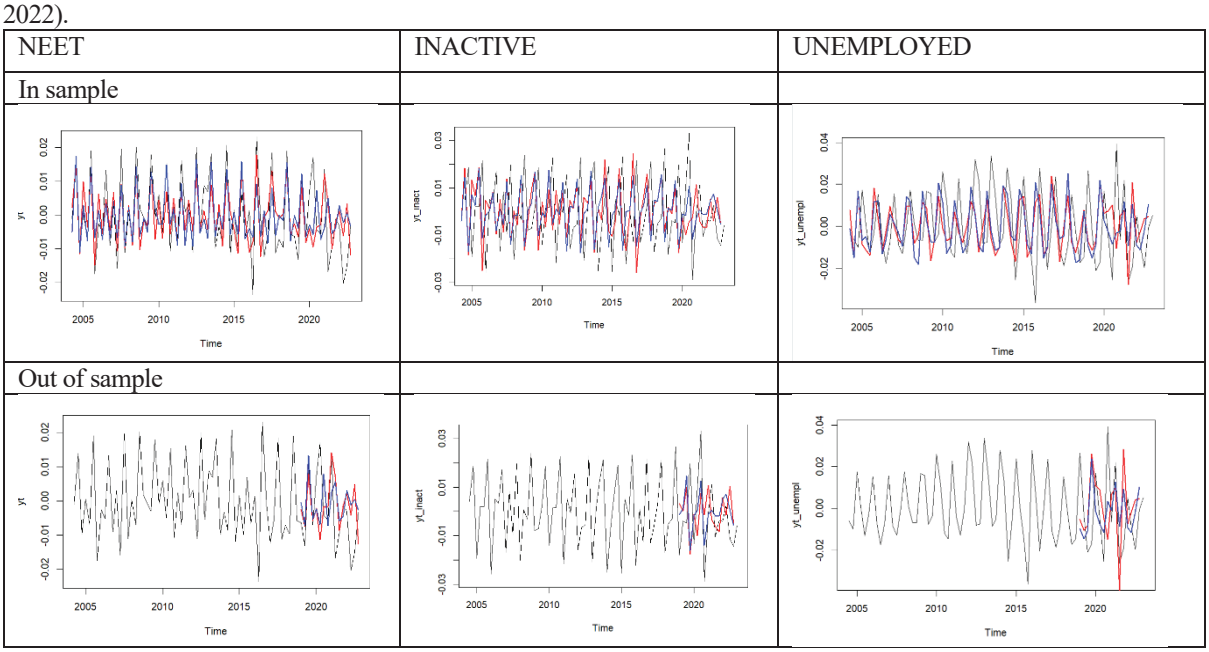
Figure 2- Key predictors of NEET, inactivity, and unemployment conditions based on random forest algorithm.

NEET	Inactive	Unemployed
		

However, based on the Random Forest algorithm (Figure 2), the most important determinant for the inactive and the unemployment rates is the share of temporary unemployment, while for the

NEETs, it is just surpassed by GDP. The share of tertiary educated, very relevant for both inactive and unemployed, is not among the most critical factors for the NEET rates, probably because of the different effect it has on these indicators, that is, increasing the share of unemployed, but decreasing the inactivity rate.

Figure 3 – In-sample and out-of-sample ability of fitting. Period selected for the forecast: last 4 years (from 2019 to 2022).



Black: empirical indicator; red: regressive model; blue: random forest.

Table 3 – Goodness of fit indicators.

	NEET		Inactive		Unemployed	
	Regression	RF	Regression	RF	Regression	RF
In sample						
RSME	0.0080	0.0080	0.0098	**0.0095**	0.0121	**0.0118**
MAE	0.0062	**0.0060**	0.0074	**0.0070**	0.0095	**0.0086**
Out of sample						
RSME	0.0120	**0.0116**	0.0164	**0.0158**	0.0201	**0.0185**
MAE	**0.0083**	0.0090	**0.0106**	0.0112	0.0157	**0.0124**

Comparing the two methodological approaches, we can see from Table 3 that they have a good and similar performance. However, Random Forest shows the best performance in the in-sample estimates in almost all the cases. Conversely, when we compare their forecast capacity, Random Forest shows a better prediction capacity for all three indicators according to the RMSE. In contrast, according to MAE, the traditional regression model shows a better capacity for the NEET and inactivity rates.

4. Conclusions

This paper analyzes the NEET rates and their determinants in Italy from 2004 to 2022. The novelty of our approach goes at least in two directions. First, besides identifying the main determinants of the NEET rates, we add to this analysis the estimation of two other important indicators: inactivity and unemployment rates. The age class chosen for this analysis is the same for the three indicators: 15-34 years. As public data were not available with a quarterly frequency for this age class, we reconstructed these time series using *ad hoc* elaboration from the LFS. Estimating these three different models allowed us to identify similarities and differences among the three indicators. This is very important, as the NEET rates include both inactivity and unemployment. Second, we adopt two different methodological approaches: the traditional regression model for

time series based on the delayed dependent variables and Random Forest, a new machine learning algorithm. We compare their ability to fit the time series overall and consider the final period of 4 years (16 observations) to make and assess their forecast capacity. The period chosen for forecasts is very challenging, as it involves the years from 2019 to 2022, characterized by the socio-economic and unprecedented shock due to the COVID-19 pandemic.

Results highlight that the same determinants affect the three indicators, but some peculiarities can be identified. While the reduction of the NEET and inactivity rates is mainly associated with an increase in the GDP, it is more important to increase labour productivity to reduce unemployment. Further, while the increase in the share of tertiary educated acts reducing the share of inactive, for the unemployment rates, the contrary happens. This is probably because a high level of education increases the motivation to search for a job besides falling in discouragement and, consequently, renouncing the search. Finally, comparing the two different methodological approaches, while their forecast capacity is much similar, differences in their capacity of investigation may be identified. The regression analysis allows us to precisely determine the response of the outcome variable to variations in the covariate values; conversely, RF allows ranking the covariates according to their capacity of influencing the outcome variable. We can then conclude that these two techniques complement each other rather than being in competition.

References

Alfieri, S., Rosina, A., Sironi, E., Marta, E., Marzana, D. (2015). Who are Italian 'NEETs'? Trust in institutions, political engagement, willingness to be activated and attitudes toward the future in a group at risk for social exclusion. *Rivista Internazionale di Scienze Sociali*, **123**(3), pp. 285–306.

Berigel, M., Boztaş, G. D., Rocca, A., Neagu, G. (2024). Using machine learning for NEETs and sustainability studies: Determining the best machine learning algorithms. *Socio-Economic Planning Sciences*, 94, pp. 1019-21.

De Luca, G., Mazzocchi, P., Quintano, C. Rocca, A. (2019). Italian NEETs in 2005–2016: Have the recent labour market reforms produced any effect?, *CESifo Economic Studies*, **65**(2), pp. 154–176, https://doi.org/10.1093/cesifo/ifz004.

De Luca, G., Mazzocchi, P., Quintano, C., Rocca, A (2020). Going behind the high rates of NEETs in Italy and Spain: The role of early school leavers. *Social Indicators Research*, 151, pp. 345–363, https://doi.org/10.1007/s11205-020-02370-3.

Mascherini, M. (2019). Origins and future of the concept of NEETs in the European policy agenda. J.O'Reilly (ed.) *Youth Labor in Transition*, pp. 503- 529.

Pastore, F., Quintano, C., Rocca, A. (2021). Stuck at a crossroads? The duration of the Italian school-to-work transition, Inter*national Journal of Manpower*, **42**(3), pp. 442-469. https://doi.org/10.1108/IJM-05-2020-0199.

Pastore, F., Quintano, C., Rocca, A. (2022). The duration of the school-to-work transition in Italy and in other European countries: A flexible baseline hazard interpretation, *International Journal of Manpower*, **43**(7), pp. 1579-1600, https://doi.org/10.1108/IJM-03-2021-0135.

Quintano, C., Mazzocchi, P., Rocca, A. (2018). The determinants of Italian NEETs and the effects of the economic crisis. *GENUS: Journal of Population Sciences*, **74**(5), pp. 2–24.

Rahmani, H., Groot, W., Rahmani, A. M. (2024). Unravelling the NEET phenomenon: a systematic literature review and meta analysis of risk factors for youth not in education, employment or training. *International Journal of Adolescence and Youth*, 29(1), 2331576.

doi.org/10.26398/asaproc.00104

AI and geostatistics: Innovative solutions for reducing and storing CO2

Angela Maria Digrandi[a], Pasquale Cimmino[b]
[a] CNR, Iriss Senior Associate, Naples, Italy, [b] Istat, Former Research Technologist, Naples, Italy

1. Introduction

By leveraging large datasets, artificial intelligence (AI) can enhance the sustainability and efficiency of cities in response to climate change, addressing emerging needs and challenges. AI-driven "smart cities" can minimize resource waste by optimizing water and energy usage. However, in light of the broader climate emergency, smart technologies should also be applied to peripheral urban areas, agricultural regions, and forests. Artificial intelligence is used to develop predictive models that help determine the best land uses based on environmental conditions and socio-economic needs, thereby minimizing environmental impact and maximizing land-use efficiency.

2. AI systems for land use and climate change monitoring at the European level

In the 1980s, the European Commission recognized the necessity for a harmonized land cover dataset across Europe, which led to the creation of the Corine (Coordination of Information on the Environment) program. The first Corine Land Cover dataset was produced in 1990, serving as a cornerstone of the Copernicus Land Monitoring service. The current Corine Land Cover (CLC) provides a detailed inventory of land cover and land use in Europe, with 44 thematic classes that are updated every six years. The Corine Land Use program integrates advanced AI technologies into its land management processes. By utilizing machine learning algorithms and neural networks, AI can analyze extensive geospatial, climatic, and demographic data to optimize resource use and promote sustainable practices. This enables real-time monitoring of land changes, allowing for proactive and sustainable interventions by quickly identifying potential risks, such as soil erosion or habitat loss.
Satellite data provide broad spatial coverage and geographical information; however, they lack socio-economic details that geostatistics offers through censuses and surveys. The pre-processing of data is aimed at combining satellite and statistical information, keeping in mind that the automatic analysis of satellite images requires advanced algorithms for object recognition, while geostatistics has a disadvantage regarding spatial coverage and timeliness.

This study employs an approach that integrates data from the Copernicus-Corine system (implemented through AI) with official statistical data from Istat, Ispra, and Arpac. An economic assessment methodology for the forest system regarding carbon sequestration was applied; this methodology feeds into tools such as Payments for Ecosystem Services (PES), which are essential for valuing ecosystem services. The document proposes a methodology for economically assessing forest CO_2 sequestration, applied to the Campania Region, providing an integrated framework to address climate change using statistical indicators alongside advanced methodological approaches. Experimentally, the analysis scope is expanded to olive cultivation.

3. The integration process of Corine CL with official statistical data

Integrating AI into Corine Land Use operations is a key driver of innovation and sustainability in land management, analyzing geospatial and climatic data with advanced algorithms. Among the statistical methods used, AI's application to classifications stands out, allowing artificial intelligence to organize and categorize data automatically, facilitating the understanding and analysis of large territorial datasets. In AI and statistics, "classification" refers to a methodology in which a predictive model assigns a label or category to new data based on what it learned from previously labelled data.

Starting from satellite images and geospatial data, the classification model - trained on labelled data - examines the features of each geographic point (such as colour, texture of images, or coordinates) and determines which category that point belongs to. The data preprocessing operation aimed to spatially align satellite and statistical data by georeferencing them, cropping the satellite data, and normalizing the geostatistical data (Figures 1-2-3-4). The quality of the linkage between satellite and statistical data was verified through geostatistical simulations.

Figure 1- Corine LC, European



Figure 2-Corine LC, Global Dynamic Land Cover

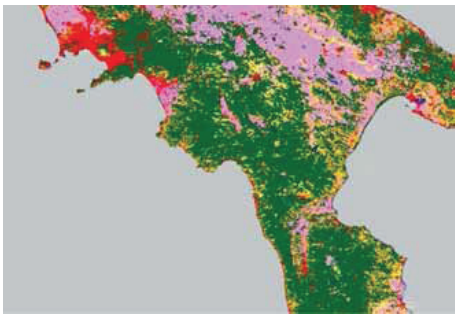


Figure 3 – Corine LC, deciduous and coniferous forests

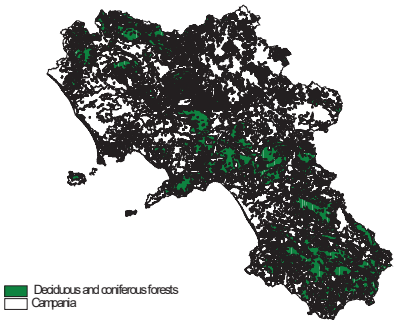
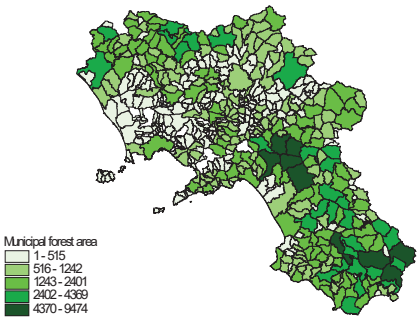


Figure 4- Istat, municipal forest areas



4. European policies to prevent greenwashing through artificial intelligence and for agricultural subsidies

The regulatory framework is the EU's unified structure for sustainable development and reporting, providing a solid foundation for defining standards based on a coherent and comprehensive architecture that include various reporting levels and sustainability topics categorized under ESG+[1].

EU Regulation 2024/1689 governs AI, emphasizing economic, environmental, and social benefits of these technologies. AI can improve forecasts, optimize operations and resources, and customize digital solutions with positive impacts in agriculture, healthcare, food security, education, and infrastructure management. The regulation emphasizes the significance of AI literacy to enable informed decision-making and ensure respect for fundamental rights; it promotes innovation while ensuring the quality and integrity of data used to train AI models. The regulation includes measures for classifying high-risk AI models and requires providers of these systems to maintain high-quality

[1] https://www.consilium.europa.eu/it/press/press-releases/2024/02/05/environmental-social-and-governance-esg-ratings-council-and-parliament-reach-agreement/

management standards. It also encourages the development of AI solutions that offer social and environmental benefits and supports SMEs and start-ups in accessing resources and regulatory experimentation.

The recently approved European regulations and the Italian applications established by Crea and Ispra aim to define standards and practices for beneficiaries of agricultural sustainability aids. These regulations include calculating quantitative indicators based on business statistics to monitor actions for biodiversity protection and environmental sustainability. Data collection and harmonization are particularly complex due to crop diversity, climatic variables, and fragmentation within Italian agricultural infrastructure, which is characterized by various uses and farm sizes. Challenges also include a scarcity of timely data on the effects of agricultural investments and soil changes caused by adverse climate events such as landslides, fires, and increasing river salinity. In this context, the support provided by satellite surveys, such as those offered by the European Copernicus system, becomes crucial for agricultural, forestry as well as biodiversity conservation policies; applied statistics play an increasingly important role in data analysis by integrating information on land use along with climatic conditions while accounting for extreme events, and related damage from radar and satellite sources with administrative sources.

The new Common Agricultural Policy (CAP) offers further research incentives by introducing "eco-schemes" which represent 25% of the total budget and encouraging sustainable techniques within agriculture, including olive cultivation. The relevant eco-scheme for olive growers is no. 3 "Safeguarding Olive Trees of Particular Landscape Value", aimed at conserving olive groves with landscape value; requirements include biennial pruning alongside a ban on burning residues, and it supports financing to improve competitiveness and sustainability, including olive pruning.

5. Case study: construction of a synthetic performance indicator for CO2 sequestration in agriculture in subsections of ecoregions

After establishing the linkage between satellite data and official statistics, a composite indicator was developed to measure the potential of forests and olive groves in CO_2 storage. Indicators were selected through an iterative process, incorporating relevant variables and focusing on the reuse of pruning by-products. Redundant indicators were eliminated to simplify the understanding of the results. The authors started with their previous research on hydroclimatic risk within Ecoregions and, based on the aforementioned European regulations, developed a project measuring potential forest ecosystems in absorbing CO_2 for risk reduction (Fig.5: hydroclimatic risk). They expanded the analysis toward the less-studied contribution of tree crops. In addition to calculating economic value associated with, CO_2 sequestration applicable within payment systems (PES), the study expanded its focus toward the entire life cycle of olive crops, analyzing waste and pruning utilized by farms, focusing on protected areas while discriminating results at the subsection level within the Ecoregions (Fig.6).

Figure 5 - Synthetic hydroclimatic risk

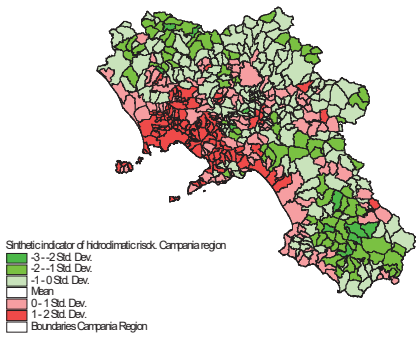


Figure 6 - Broadleaf forests and olive groves

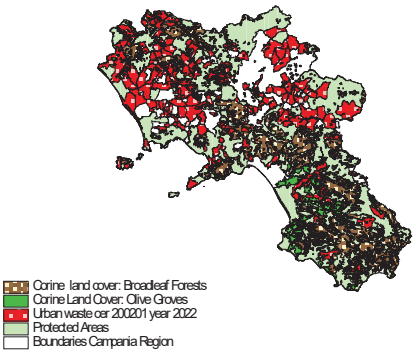


Figure 7- Broadleaf forests in protected areas by Ecoregion subsection

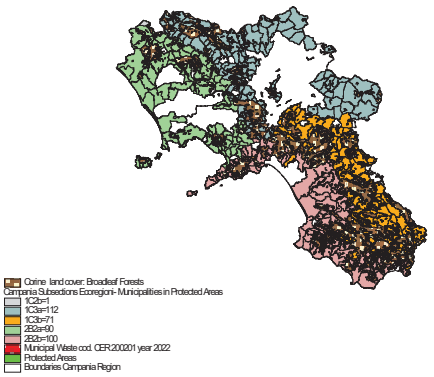
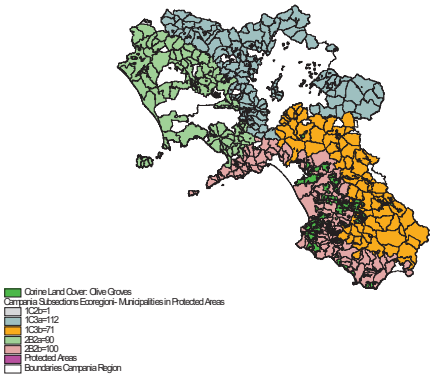


Figure 8- Olive groves in protected areas by ecoregion sub



Source: author's elaboration on Corine LC and Istat data

The geostatistical analysis, developed using the classification system of Ecoregions (Fig. 7) and satellite data processed by Copernicus-CLC, allows for the specific localization of tree species, both forest types by macro classification and fruit tree crops, with a particular focus on the localization of olive groves (fig.8). These are specifically and predominantly found in subsection 2B2b of the Mediterranean division of Italian Ecoregions.

The construction of the synthetic index encompassed several preparatory phases, including iterative phases of reasoned selection of available base indicators and their standardization through the creation of indicators related to the resident population and the surface area of protected areas at the municipal level (Fig. 9). The synthetic indicator highlights areas with potential for increased CO_2 sequestration by forest and olive groves in protected areas, aimed at producing bioenergy (Fig. 10).

The bioenergetic potential synthetic indicator, calculated using the Mazziotta-Pareto MPI-method, measures the potential for activating biodegradable waste collection to reduce energy costs.

This method transforms elementary indicators into distances from the mean, standardized by the standard deviation. This process yields indicators for each observed unit, resulting in a mean of 0 and a standard deviation of 1, with increasing values. They decided to apply the negative method according to the formula outlined below.

$$z_{ij} = \begin{cases} 100 + \dfrac{(x_{ij} - M_{x_i})}{S_{x_i}} 10 & \text{For indicators with increasing direction} \\ 100 - \dfrac{(x_{ij} - M_{x_i})}{S_{x_i}} 10 & \text{For indicators with decreasing direction} \end{cases}$$

A penalty coefficient is applied to the basic indicators that depend, for each territorial unit, solely on the variability of the indicators relative to the mean value. No additional weights have been applied to the indicators since the research objective is exploratory, allowing a focus on the most significant variables for which it is possible to reconstruct their contribution to the building of strategy policies.

Figure 9-Linkage between satellite data and statistical data collected on waste collection CER 2001201

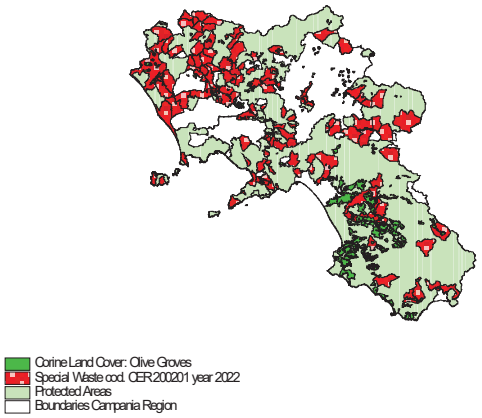
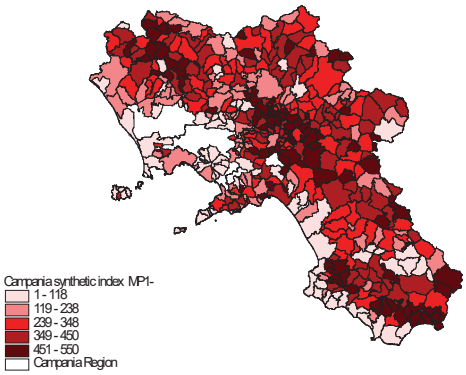


Figure 10- Synthetic indicator of bioenergy potential in protected areas



Source: author's elaboration on Corine LC, Arpac, Ispra, and Istat

6. Economic valuation of forests and historical olive groves

The geostatistical analyses developed in the previous sections pave the way for exercises in the economic evaluation of natural ecosystems and their potential in rebalancing, aligning with urban regeneration policies. They also provide simulation spaces for a new economic policy capable of integrating two visions of the future: the strictly naturalistic and ecological one, alongside a more technological perspective. This leverages artificial intelligence as an innovative area of data analysis and a place of interaction between technological and humanistic intelligence.

As seen in Section 3, AI can integrate innovative technologies and support scientific research in the fight against climate change, can highlight the efforts of territories and businesses adopting eco-sustainable production systems. This contributes significantly to ensuring that, as European regulations require, public funds are not used for greenwashing activities. AI can also support an economic policy incorporating corporate accounting values of cooperation, innovation, and education, surpassing evaluations based solely on investments or economic losses associated with catastrophic events by accounting for the direct contributions of innovation and risk prevention.

The most recent European Emissions Allowance Auction Report (2022) estimates the average price of CO_2 sequestration at 82 euros per tonne. Considering that the forestry heritage in Campania, covering about 380,000 hectares, absorbs 2,172,800 tonnes of CO_2 annually, the economic market value is approximately 17,172560 euros[2].

It should be emphasized that the potential hedonic value of forest landscapes and historical olive groves characterizing Southern Italy and the entire Mediterranean basin is significant. Consequently, calculating production loss in unfavourable years due to adverse climate events, or olive tree epidemics must also consider replanting costs and loss of landscape value. Shadow pricing related to the proximity value to historical natural landscapes in real estate transactions should also be included, as well as the economic value of slow tourism, ranging from rural tourism to historical trails, which can be severely affected by the destruction of forests, olive groves, vineyards, and historic orchards. The following are the main value indicators for the natural landscape of historical olive groves which will be developed within the research in further phases: tourism value; property value increase (ancient olive trees can increase the value of surrounding real estate properties), CO_2 capture and

[2] Nestico' et al. 2023

sequestration, soil erosion costs; biodiversity loss (the biodiversity lost due to removing olive trees can reduce agricultural resilience and increase vulnerability to diseases and pests).

In this context, synthetic indicators are crucial for economic evaluation and continuous monitoring of technological and social effects.

7. Conclusion

The use of artificial intelligence (AI) in territorial planning enhances decision-making and reduces environmental impact by enabling real-time soil monitoring for sustainable management based on updated data. This study examines the reutilization of agricultural and arboreal waste in the ecoregions of Campania, combining geoclimatic and satellite data to build a synthetic indicator of "Ecosystem potential" of forested and olive-growing areas. The construction of the synthetic index for olive-growing areas, even small ones, involves agricultural companies and biodegradable waste collection firms. The synthetic index uses an approach that facilitates the interpretation of results and the identification of opportunities for implementing policy strategies. Furthermore, it promotes research centres and startups that utilize such waste to produce biogas with minimal CO_2 generated during the transformation process.

The study presented is the first phase of research aiming to expand Payment for Ecosystem Services (PES) usage of forest data in protected areas and to measure the hedonic value of olive landscapes, using shadow prices to assess the economic impact of these landscapes on real estate prices by isolating the landscape's contribution to property value while considering size and location. Thanks to their aesthetic and cultural value, forest and historical olive landscapes can significantly increase the value of surrounding properties and reveals a new role for public entities and agricultural enterprises.

References

Blasi, C., Capotorti G., Copiz R., Guida, D., Mollo, B., Smiraglia, D., Zavattero, L.(2018). *Terrestrial Ecoregions of Italy. Map and Explanatory Notes.* Global Map srl, Firenze.

EU-https://www.consilium.europa.eu/it/press/press-releases/2024/02/05/environmental-social- and-governance-esg-ratings-council-and-parliament-reach-agreement/

Crea: https://creafuturo.crea.gov.it/3922/

Digrandi, A.M, Cimmino P., Monteleone, G. (2023). Geostatistics for the analysis of complex Ecosystem., *Italian Journal of Applied Statistics*, 35(3), supplement.

European Energy Exchange-https://www.eex.com/en/markets/environmental-markets/eu-ets-auctions.

Ispra (2023)- https://www.isprambiente.gov.it/it/attivita/suolo-e-territorio/suolo/il-consumo-di suolo.

Istat (2023) -https://www.istat.it/non-categorizzato/classificazione-dei-comuni-secondo-le-ecoregioni-ditalia/.

Istat (2024) – https ://esploradati.istat.it/databrowser/#/it/censimentoagricoltura.

Istat - https://i.ranker.istat.it/

Mazziotta, M., Pareto, A., (2020). *Gli indici sintetic*i. Torino: Giappichelli.

Nesticò, A., Russo. R., Maselli, G., (2023). Forest ecosystem services: economic evaluation of carbon sequestration on a large scale*, Siev, Valori e Valutazioni*, pp. 17-30.

UE-https://eur-lex.europa.eu/legal-content/it/ALL/?uri=CELEX%3A32019R2088

Wunder, S., (2015). Revisiting the concept of payments for ecosystem services. *Ecological Economics,* 117, pp. 234-243.

doi.org/10.26398/asaproc.00105

Leveraging Industry 5.0 for official business statistics innovation

Diego Distefano, Paola Bosso, Giovanni Gualberto Di Paolo, Pasquale Papa
ISTAT - Italian National Statistical Institute, Rome, Italy

1. Background

The development of new technologies has started, in the last decade, a radical transformation of industrial processes, at a national and international level. Specifically, this development has led to the emergence of digitalisation paradigms such as Industry 4.0 (Xu et al., 2021), Industry 5.0 (I4.0 and I5.0), and Smart Manufacturing. The main difference between Smart Manufacturing and Industry 4.0 is mainly geographical. The Industry 4.0 project, in fact, is of European origin, has Germany among the leaders and has a particular impact in Europe. Smart Manufacturing has greater traction in the United States, thanks to an initiative called Smart Manufacturing Leadership Coalition (SMLC), a non-profit organization made up of representatives of the manufacturing supply chain, universities and research centers. The two projects have many elements in common: interoperability between systems within the entire supply chain, creation of standards and use of increasingly advanced technical solutions to make processes more efficient, economical, flexible and scalable.

Digitalisation has led to the availability of wide amounts of data, enabling a data-driven approach for companies. This involves leveraging big data to inform and optimise business decisions. Data-driven companies view data management as a strategic asset.

This study aims to explore the potential of the above mentioned paradigms for Italian official statistics, assessing both benefits and challenges. This analysis also seeks to develop a strategy for integrating new data sources into Italian official statistics (Bender et al., 2022). A key component of this strategy involves assessing the digital maturity of Italian businesses, especially SMEs (Small or Medium Enterprises), to determine the feasibility of innovative solutions.

2. Analysis of data

The official surveys conducted by Istat, as part of the National Statistical Program, impose a variable burden depending on the size and industry of the companies. They include forty-four surveys, of which twenty-two are short-term (covering 127,482 companies), sixteen are annual, and eight are multi-year or occasional (involving approximately 395,453 companies), for a total of 452,322 companies involved[1]. As shown in Table 1, 99% of companies participating in official business surveys are SMEs (97% for short-term surveys). However, large companies have a much higher economic weight, accounting for about 23% of employment and 35% of added value.

Businesses involved in statistical surveys participate in a minimum of one survey and a maximum of 21, with many surveys repeated multiple times throughout the year, especially for weekly, monthly, and quarterly frequencies. Average survey involvement varies by size: micro and small businesses participate in an average of 1.5 surveys, while larger businesses participate in 11.8 surveys on average. Even medium-sized companies have a relatively high average involvement of 6.2 surveys.

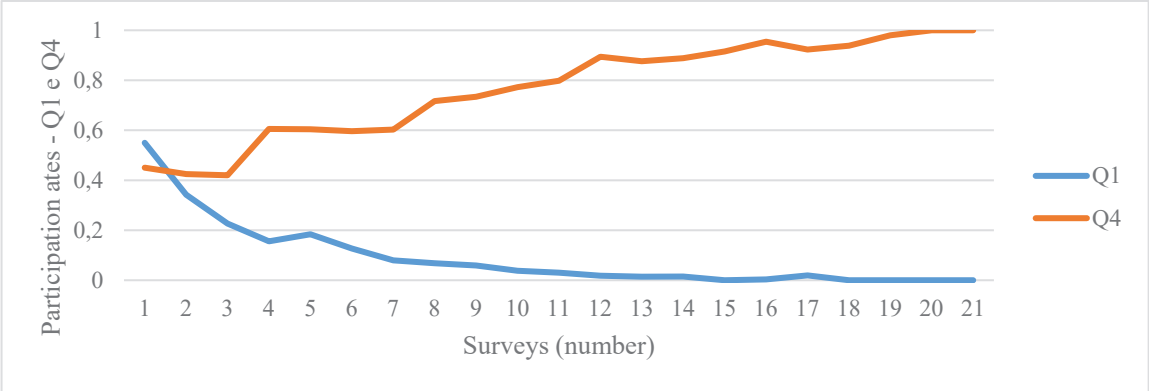
[1] The reference is to the last completed edition of each direct survey. All surveys are conducted using the CAWI (Computer-Assisted Web Interviewing) tecnique.

Table 1. Businesses involved in ISTAT official surveys, by size range and type of direct survey.

	Type of direct survey					
Businesses size range	*All*	*%*	*Structural*	*%*	*Short-term*	*%*
Micro and small	430,449	95	373,900	95	111,420	87
Medium-sized	18,425	4	18,135	4	13,036	10
Large	3,448	1	3,418	1	3,026	2
Total	**452,322**	**100**	**395,453**	**100**	**127,482**	**100**

For each business involved in one or more surveys, participation rate (*r*) is calculated as the ratio of surveys completed to surveys requested. The first (Q1) and fiurth (Q4) quartiles are calculated in relation to the number of surveys. As an example, companies involved in a single survey have a first quartile of approximately 0.6, while those involved in more than 10 surveys have a first quartile below 0.1 (Figure 1). Analyzing the first and fourth quartiles of participation rates reveals a direct relationship between participation and the number of surveys. This result is likely influenced by the company size, as larger companies tend to participate in more surveys.

Figure 1. Participation rates of businesses in Istat surveys by number of surveys, Q1 and Q4 quartiles.



It is noteworthy that 84% of large enterprises fall into Q4, indicating a particular "loyalty" of large enterprises in participating in official surveys. In contrast, micro enterprises (which are typically involved in 1.5 surveys) show a near perfect balance between responding and not responding. Specifically, 53% of micro enterprises fall into Q1, while 43% are placed in Q4.

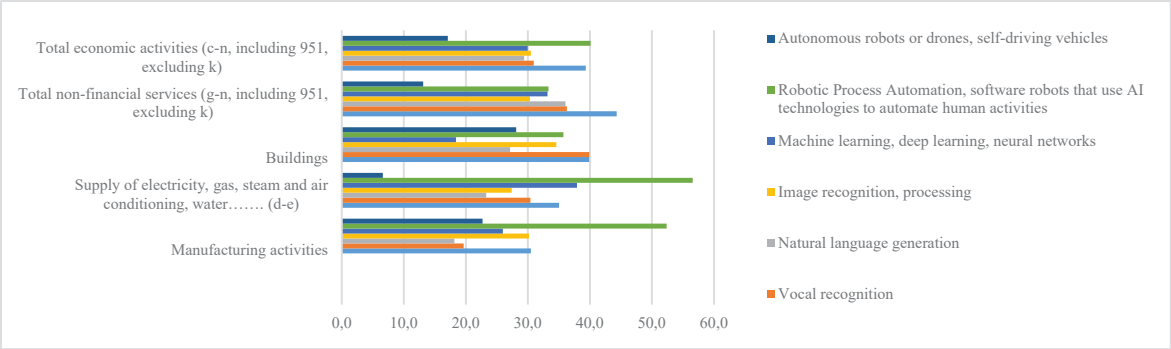
Official direct surveys on businesses are becoming increasingly difficult to conduct. Respondents are becoming more reluctant to participate, and statistical agencies are struggling to afford the costs of conducting surveys. These challenges are exacerbated by the need to meet high quality standards for survey outputs. In this context, the digitalisation process and the development of the fourth and fifth industrial revolutions (I4.0 and I5.0) offer new opportunities to improve the efficiency and effectiveness of surveys (De Waal et al., 2020).

3. Digitalisation degree of Italian companies

Results of the annual survey on diffusion of information and communication technologies in businesses (in short ICT survey) provides valuable data on the digitalisation of Italian businesses, including SMEs. The 2023 survey found that only 22.2% of companies have a high or very high level of digitalisation, but this figure rises to 68.2% for larger businesses with 250 or more employees. The level of digitalisation varies significantly by dimension, with small businesses lagging behind at 19.3% and medium-sized businesses at 47.0%. ERP (Enterprise Resource Planning) software is used by 42.7% of companies, and business intelligence by 17.6%. 61.4% of companies use cloud computing services, including 10.9% for advanced ERP (13.1% for manufacturing).

The survey highlights the adoption of AI too. While only 5.0% of businesses overall use AI, this figure rises to 24.1% for large companies. Furthermore, a relevant portion of AI-using businesses in the manufacturing sector (52.5%) apply AI to their production processes, suggesting a promising trend for the future of manufacturing. In manufacturing AI adoption isn't only for large companies (46.1%): 38.3% of small businesses (10-49 employees) also use it. The main barrier in using AI is a lack of skills (55.1%, rising to 58.1% in manufacturing). In the manufacturing sector, AI is primarily used for robotic process automation and software automation (Figure 2).

Figure 2. Per cent of businesses using AI software or systems, by purpose and sector of economic activity



Source: Istat – ICT survey, year 2023.

Overall, the survey paints a picture of advanced digitalisation among larger companies, although there are gaps and delays in certain areas such as AI. The level of digitalisation is not uniformly distributed among all businesses. In this regard, the situation is much more different for SMEs, with significant differences emerging across sectors, types of technology, and areas of application. Data from the ICT survey points out that businesses with a high degree of digitalisation represent a relevant portion of Italian SMEs, demonstrating a "readiness" for the implementation of new technologies, particularly in the manufacturing sector. The diffusion of ERP management systems, even advanced ones, is notable even among small-sized companies. A great and a significant further impulse to the development of these activities may derive from the incentive policies for the 4.0 and 5.0 transition promoted and put into action by national governments. On this regard, in Italy, the Operational Circular No. 25877 of August 16, 2024, on Transition 5.0 provides the Ministry of Enterprises and Made in Italy guidelines for the implementation of the Transition 5.0 Plan, offering technical clarifications useful for the correct application of the new incentive regulations.

4. A new strategy to official business survey

4.1 Basic definitions

NIST (National Institute of Standards and Technology) defines smart manufacturing as: “fully integrated and collaborative manufacturing systems that respond in real time to meet changing demands and conditions in the factory, supply network, and customer needs“. Smart manufacturing uses cloud technology, combines human creativity with machines and AI, and aims for faster, more precise, and personalized production with less waste and more flexibility (Rüßmann et al., 2015). Very similar to the I4.0 paradigm definition, that involves establishing a highly automated and interconnected production environment, in which devices, machines, people, and systems can communicate and exchange information with each other in real time.

The concept of I4.0 has evolved into I5.0, which encompasses not only the technological dimension but also the human, environmental, and resilience issues. This broader framework recognizes the importance of integrating these elements to create a more sustainable and

resilient industrial future (Breque et al., 2021). In both I4.0 and I5.0, technological advancements drive a data-centric approach, where data management becomes a strategic business imperative.

Embracing a data-driven approach entails leveraging the availability of big data within businesses and effectively integrating it into the decision-making process. Every component of the production chain, including products, is interconnected through the integration of sensors, measurement, and monitoring instruments, such as RFID chips. This interconnectedness extends to the supply chain, which exists simultaneously in both the physical and virtual realms.

4.2. Operational procedure for leveraging new data sources for statistical survey processes

A targeted operational procedure aimed at exploiting the potential offered by I4.0 and I5.0 technologies for official statistical purposes fits the collection of statistical data on businesses within a multisource approach (Salemink et al., 2020; Saraiva Dos Santos, 2022). Specifically, the primary goal is to conduct a thorough analysis of the operational opportunities presented by I4.0 and I5.0 technologies in the realm of official statistics. It is carried out by a multilevel approach, examining impacts across different investigation areas, statistical, technological, regulatory, organisational. The activity involves the implementation of some preliminary and operational steps, summarized as follows:

- Preliminary step 1: Analysis of the basic characteristics and requirements of the current official business survey system, mainly based on direct surveys, with specific regard to burden of respondents and costs. In terms of costs, business surveys have lower direct expenses compared to those targeting other types of survey units, as they are predominantly carried out using the CAWI method. However, they still demand significant internal resources and incur additional costs for support services such as contact centers for assistance and reminders, translations, questionnaires design, set-up and optimization, and more.

- Preliminary step 2: Analysis of the degree of diffusion, also through the use of proxy variables, of the I4.0 and I5.0 technologies in Italian businesses, in order to draw indications on the "coverage" and the possibilities of extending the solutions identified. To properly assess the resulting opportunities and challenges, detailed insights into the level of digitalisation among Italian businesses are crucial, as briefly noted in the ICT survey results mentioned in Section 3.

- Preliminary step 3: Case study on retail sales survey: The initial steps towards the application of the methodology outlined have been undertaken for the monthly retail sales survey. This is a monthly short-term survey that collects data on sales and point of sales. The variables for the survey were identified as those potentially available[2] within 5.0 management platforms, accompanied by their corresponding definitions. It is crucial to clarify these variables in order to understand their alignment with the information available in the 5.0 systems, as well as any discrepancies that will necessitate the application of transformation rules. Following a thorough analysis of the functionalities and characteristics of various 5.0 platform types available on the market, a selection was focused on the most suited to the project's objectives, specifically advanced ERP and manufacturing execution systems. Additionally, potential providers were identified, consisting of 3-4 companies offering ERP and manufacturing execution systems solutions.

- Operational step 1: Competence centres (CSs) are established by the Ministry of Enterprises and Made in Italy with the aim of providing services and assistance to companies in 5.0 transition. A step in order to exploit the potential offered by I4.0 and I5.0 technologies

[2] Example of variables of interest for retail sales survey:
- total sales value by product groups and point of sale type;
- total number of the company's points of sale;
- number of opening days in the month per point of sale;
- total sales area per point of sale (sqm);
- number of employees per point of sale;
- sales value by channels: internet, mail order, other.

for official statistical consists of establishing synergies, information exchange and collaboration with the CSs. The goal is to delve into the digitalisation incentives introduced by the Italian government, establish connections with businesses that have benefited from competence center services.

- Operational step 2: Integrated management platforms are a crucial tool for official statistics to harness the large volumes of data generated by business digitalisation processes. Analyzing the data categories provided by leading advanced and integrated management platforms (second-generation ERP) is a necessary step in fully utilizing new technologies. For this reason, Istat has identified a set of key integrated platform providers to initiate contact and conduct structured interviews;

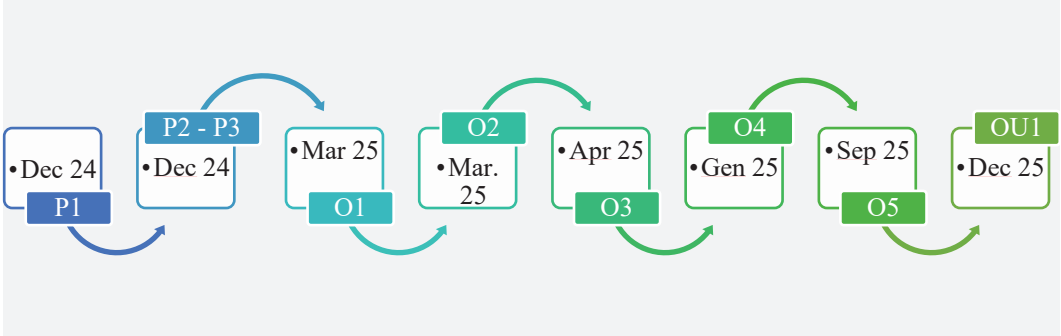
- Operational step 3: Evaluation of acceptability of the new approach among a purposive sample of respondents; A data collection technique based on automated M2M (Machine-To-Machine) methods is more "invasive" compared to direct surveys carried out via web questionnaires (Snijkers, 2022). A key step in understanding the potential of this new data source is the evaluation of attitudes, specifically the acceptability of the method by data providers. For this purpose, a key step in the process is identifying a small, targeted sample of businesses with different structural characteristics (size, macrosector, region) and varying levels of digitalisation.

- Operational step 4: Detailed maching of data produced by the advanced ERP platforms, with statistical information needs. The challenge is to determine to what extent the information produced by management platforms can be used directly or with necessary transformations and processing for official statistical outputs, as required by EU regulations (particularly EU Regulation 2019/2152 on European business statistics) and national dissemination channels (press releases and detailed reports).

- Operational step 5: The final operational step involves realizing an experimental parallel data transmission by a selected sample of highly digitized companies, in order to compare it with traditional transmission methods and verify compliance with the content and quality requirements set by official statistics.

- Output 1: The objective of the operational procedure is to design a range of solutions that, by leveraging the potential of various platforms, enable the establishment of a set of M2M transmission options that adapt to the different organizational, structural, and technological characteristics of statistical data providers.

Figure 3. Deadlines set for each step of the process.



5. Discussion

The digitalisation of production chains generates wide amounts of data, offering new perspectives for official business statistics. The challenge lies in transforming business data into official statistics. The leverage of businesses digitalisation offers possible advantages for official statistics, including a substantial reduction in the burden on businesses, the potential for reducing the costs of direct surveys in the medium term, and the opportunity to improve the

quality of statistical outputs in terms of timeliness and other components of the total survey error. Government incentives for the 4.0 and 5.0 transition can provide an additional boost. Moreover, the digitalisation of production processes is increasingly becoming a competitiveness factor that will significantly influence Italian businesses.

Italian official statistics must equip itself to enhance its capabilities and expand its data sources. On the other hand, the operational use of new data sources presents some challenges, primarily related to the initial costs of engineering the operational use of these new sources. While Industry 4.0 and 5.0 paradigms are increasingly adopted in Italian production, coverage is still partial, especially for SMEs.

The initial phase requires investments in generalized technological solutions for the direct acquisition of data, especially considering the heterogeneity of management platforms, as well as statistical and organizational solutions to maintain parallel traditional direct data collection processes, those from administrative sources, and those from M2M sources. Other challenges relate to managing potential resistance from businesses involved in the survey to accept direct and automatic data transmission methods linked to their management systems and possibly their databases.

Finally, challenges remain regarding the consistency between definitions and classifications adopted by ISTAT and those adopted by data providers, which may require ad hoc transcoding and transformation operations. In conclusion, the short-term benefits are not immediate and do not outweigh the challenges, highlighting the need for further investigation into the topic. As with all innovations, the initial phase always represents a gamble and requires additional investment and resources to realize the medium-to-long term potential for benefits and efficiency gains.

References

Bender, S., Sakshaug, J.W. (2022). Data sources for business statistics: What has changed?. *The Survey Statistician*, 85, pp. 10–18.

Breque, M., De Nul, L., Petridis, A. (2021). *Industry 5.0: Towards a Sustainable Human-Centric and Resilient European Industry,* Brussels, Belgium: European Commission.

De Waal, T., Van Delden, A., Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. *International Statistical Review,* 88(1), pp. 203-228.

Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., Harnisch, M. (2015). *Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries.* Boston Consulting Group.

Salemink, I., Dufour, S., Van der Steen, M. (2020). A vision on future advanced data collection, *Statistical Journal of the IAOS,* 36, pp. 685–699.

Saraiva Dos Santos, P. (2022). Organizational responses to multiple data collection - Administrative data unit, *UNECE Expert Meeting on Statistical Data Collection, Towards to a New Normal? , 26 - 28 October 2022, Rome, Italy*.

Snijkers, G. (2022) System-to-system data collection in business surveys applied to an agricultural survey: A proof of concept, *UNECE Expert Meeting on Statistical Data Collection, Towards to a New Normal?* , 26 - 28 October 2022, Rome, Italy

Xu, U., Lu, Y., Vogel-Heuser, B., Wang, L. (2021). Industry 4.0 and Industry 5.0—Inception, conception and perception. *Journal of Manufacturing Systems,* 61, pp. 530-535.

doi.org/10.26398/asaproc.00106

Future orientation and psychological and social capital of Italian young adults

Angela Maria D'Uggento[a], Luigi Fabbris[b]
[a]Department of Economics and Finance, Aldo Moro University of Bari, Italy
[b]Tolomeo Studi e Ricerche, Padua, Italy

1. Introduction

Future time perspective (FTP) can be defined as the perceptions and expectations a person has about his or her future, specifically the extent to which he or she perceives the future as open-ended and full of opportunities, or as limited and closed in nature (Trommsdorff, 1994; Castersen et al., 1999). The future can be evaluated as fulfilling one's wishes or as being fearful and threatening. Accordingly, a person's future orientation can be more or less optimistic or pessimistic (Zelenski and Larsen, 2002).

The study of FTP is relevant to young people because those who have this disposition put more effort in learning, career planning and long-term behaviours than those who think less about possibilities regarding their future. In other words, thinking about the future is likely to influence human actions in the present. Indeed, research suggests that learning, academic performance, business, professional career, leisure, social relations, health and quality of life may be improved by a positive FTP (Trommsdorff, 1994; Carstensen et al., 1999; Lang and Carstensen, 2002; Peetsma and Van der Veen, 2011; Andre et al., 2018; Sekścińska et al., 2018; Kiani et al., 2020).

This paper aims to understand if and how the disposition of Italian youth to face their future is influenced by personality traits and social capital. Personality traits include four of the so called "big five" traits, three components of psychological capital, as well as self-confidence and creativity. The involved personality traits are described in Section 2. All traits considered in this paper being positively oriented, they will be referred to below as "psychological capital".

Social capital is the term used to describe the potential to achieve results through personal connections. With reference to a person, it can be used to describe his or her success in personal relationships and the skill of building networks both within and outside his or her family or the community in which he or she operates. In this paper, we consider social capital as the convolution of people's family bonds and individual activities in voluntary associations.

Our basic hypothesis is that both psychological and social capital may enhance an individual disposition to the future, while depression and other personal and social problems may limit it.

In the following, we present the results of a survey carried out on a sample of Italian adults aged 18-34 years to shed light on the relationship between the youth's disposition to the future and their psychological and social capitals, on the one hand, and depression, individual worries and problems, on the other, with a set of control variables in the background.

The research questions under analysis are as follows:

Q1: What aspects of Italian youth's personality correlate with their belief in a positive future for them?

Q2: Are there aspects of social capital that, in interaction with their personality traits, may influence the youth's FTP?

Q3: Are there problems that may limit the FTP of Italian youth? In particular, is there a link between depression - which accounts for a worrying 28.5% of our sample - with threats toward the future?

Q4: Does the females' propensity to a positive future differ from that of males?

The rest of the paper is organised as follows: Section 2 introduces the available data and the model for data analysis. Section 3 presents the main results of the statistical analysis of the collected data. Finally, Section 4 discusses the results with reference to the mainstream literature on FTP.

2. Data, model and methods

A CAWI (Computer Assisted Web-based Interviewing) questionnaire was administered to a convenience sample of Italian people via email and social media from June to December 2023. A total of 509 respondents completed the electronic questionnaire. In the following, we limit our analyses to a subset of 307 respondents whose age was between 18 and 34 years, which, in official statistics, is usually referred to as young adults. The obtained responses were clustered as follows.

FTP: Having a positive future time perspective. The propensity to future was surveyed through a set of eight dichotomous (True/False) items selected from the Zimbardo's psychometric test to evaluate FTP (Zimbardo and Boyd, 1999; Di Zio and Fabbris, 2023) and the responses were factor analysed. Factor analysis is a multivariate method to transform a set of correlated items into one (as is our case) o more linearly independent dimensions. The factor loadings were: 0.618; 0.434; -0.623; 0.440; -0.693; 0.629; 0.395; 0.528). A positive FTP was defined as the higher 30% of the factor scores after testing for one-dimensionality of the scores.

The surveyed *personality traits* were a selection of traits suggested by the psychometric literature. We selected four of the "big five" traits (Costa and McCrae, 1978): sociability (openness to others/selflessness), conscientiousness (perseverance/hard working), extraversion (vitality/ease), and stress/emotion control capacity, so ignoring openness to experience, as it was subrogated by other surveyed traits. Moreover, we considered three of the four aspects of "positive psychological capital" (Luthans et al., 2007): optimism, resilience (ability to bounce back after a negative event), and self-effectiveness (ability to fulfil targets). People were also asked for their self-esteem (confidence on their own competence: Rosenberg, 1985) and creativity (artistic ability/love of art). In the following, the nine personality traits are referred to in their entirety as *psychological capital*.

The variables that in our study define a person's *social capital* are the following: a) Living in a family with a partner and/or other people; b) At least one parent having a higher education degree; and c) Membership in a non-profit organisation.

In our study *individual expectations and interests* include a) Having children; b) Having a family of one's own; c) Interest in the natural environment; d) Interest in the Italian situation; and e) Interest in the global situation.

The individual problems include: a) *Full-blown depression* (dichotomous; computed using the nine-item 0÷3-mode PHQ–Patient Health Questionnaire proposed by Spitzer et al. (1999) and translated into Italian by Mazzotti et al. (2003): a score of cumulative responses larger than nine identifies the diagnosis of major depression); b) *Worries* for future community (Job opportunities; Economic wellbeing; Social and political stability; Environmental quality) and personal dimensions (Job opportunities; Economic wellbeing; Health; Having a family of one's own; Having children); respondents' worries were transformed into a single variable through factor analysis, after the responses were checked for one-dimensionality; and c) *Current fears* (Diseases; Contacting unknown people; Social and economic world crises; Natural phenomena).

The respondents (see Table 1) were predominantly female (58.6%) and had an average age of 23.3 ±4.2 years). The group of 18 to 24 years made up 67.4% of the sample at hand. Consequently, many of the respondents were students (63.2%) or employed (27.0%) and only 9.8% were in another position. Concerning social status, 36.5% of respondents had at least one parent with a higher degree. The majority of the sample (60.3%) lived in a southern Italian region; the others lived in a northern region (31.9%) and a few (7.8%) in central Italy.

The relational model adopted for the analysis of the collected data is shown in Figure 1. The hypothesis is that the dependent variable, FTP, may be influenced by both positive and negative aspects of personality, mood and social capital, while some control variables (sex and professional status) may influence both psychological and social capital and the dependent variable. A variant of the model that excludes severe depression as a predictor (Model 2) is also considered to potentially strengthen the causal role of minor psycho-social variables that could be inhibited by the dominance of depression as a predictor of FTP variability.

Table 1. Distribution parameters of positive FTP of Italian young adults, by personal and social characteristics (n=307)

Characteristic	*% FTP*	*% sample*	*Characteristic*	*% FTP*	*% sample*
Z_1:Sex (Male)	36.2	41.4	X_{17}:Resilience	38.1	64.2
Z_2:Student	30.9	63.2	X_{18}:Optimism	43.3	58.0
Z_3:Employee	42.2	27.0	X_{19}:Creativity	34.1	58.3
X_1:Higher education parents	36.6	36.5	X_{20}:Conscientiousness	35.0	73.6
X_2:Partner	32.9	47.6	X_{21}:Stress control capacity	40.2	53.4
X_3:Volunteer associations	34.8	51.5	X_{22}:Satisfying job (target)	28.0	51.1
X_4:Severe depression	13.5	29.0	X_{23}:Degree (target)	36.6	13.4
X_5:Worries high level	29.0	30.3	X_{24}:Econ. independ. (target)	27.7	44.6
X_6:Disease concern	35.1	37.1	X_{25}:Marriage (target)	34.8	15.0
X_7:Alien people concern	31.4	16.6	X_{26}:Research (target)	25.0	5.2
X_8:World crises concern	29.3	40.1	X_{27}:My house (target)	37.5	13.0
X_9:Interested climate issues	33.3	51.8	X_{28}:Land, garden (target)	44.4	2.9
X_{10}:Interested environment	33.7	67.8	X_{29}:Family affection (target)	51.5	21.5
X_{11}:Interested Italian situation	32.7	50.8	X_{30}:Travelling (target)	34.5	28.3
X_{12}:Interested world situation	35.1	56.7	X_{31}:Having a partner (interest)	38.1	61.6
X_{13}:Self confidence	38.1	64.2	X_{32}:Having children (interest)	40.1	57.7
X_{14}:Self effectiveness	40.6	64.2	X_{33}:Working abroad (target)	33.3	32.2
X_{15}:Extraversion	36.8	56.7			
X_{16}:Sociability	35.4	72.6	Total	32.6	100.0

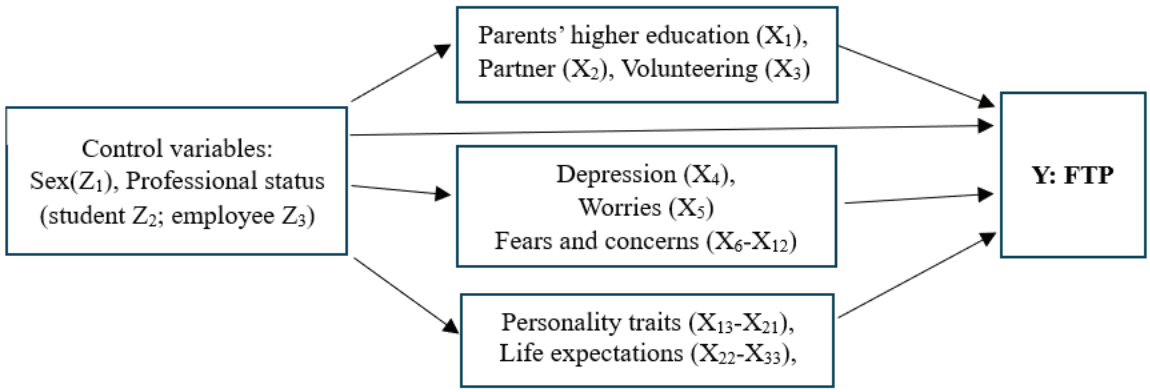


Figure 1. Blocks of variables involved in a possible causal relation with Italian young adults' FTP.

The logistic regression model for multivariate analysis is (Hosmer and Lemeshow, 2000):

$$logit\ [p(Y=1)] = \beta_0+\beta_1 X_1+\cdots+\beta_j X_i+\beta_{j+1} Z_1+\cdots+\beta_{j+3} Z_3$$

where $logit\ [p(Y=1)] = ln[(p/(1-p)]$. The *Zs* represent the control variables that were forced into the model and X_i (i = 1, …, 33) represents a predictor selected according to its significance level. A significance level < 0.10 was assumed for the selection of predictors. The statistical analyses were carried out using IBM SPSS software (IBM, 2017), Factor and Logistic Regression commands.

3. Results and discussion

The main results of the logistic regression analysis (Table 2) showed that Italian young adults' FTP is well explained by the variables hypothesised as predictors. In fact, Model 1, which includes five predictors selected through a forward stepwise technique, in addition to the forced control variables, explains 24.5% of the positive-FTP variance, while Model 2, which excludes severe

depression, includes eight predictors and explains 24.8% of the FTP variance.

Table 2. Beta estimates of logistic regression with (Model 1) and without (Model 2) severe depression as regressor of Italian young adults' FTP as criterion variable (n=307).

	Model 1				Model 2			
	β estimate	s.e.	Exp(β)	95% CI Exp(β)	β estimate	s.e.	Exp(β)	95% CI Exp(β)
Z_1: Male	0.394	0.280	1.483	0.856-2.570	0.643*	0.288	1.903	1.082-3.346
Z_2: Student	0.374	0.555	1.453	0.490-4.311	0.426	0.564	1.531	0.507-4.621
Z_3: Employed	0.571	0.588	1.770	0.560-5.601	0.774	0.586	2.168	0.687-6.841
X_4: Severe depression	-1.133**	0.359	0.322	0.159-0.651				
X_{14}: Self effectiveness	0.668*	0.330	1.950	1.020-3.725	0.764*	0.357	2.147	1.066-4.322
X_{18}: Optimism	0.674*	0.321	1.962	1.046-3.678	1.119**	0.346	3.061	1.554-6.029
X_{20}: Conscientiousness					-0.751°	0.401	0.472	0.215-1.035
X_{23}: Degree (target)					0.736°	0.409	2.088	0.937-4.652
X_{29}: Family affection (target)	0.834**	0.320	2.303	1.229-4.316	1.303***	0.348	3.682	1.862-7.281
X_{30}: Travelling (target)					0.672*	0.326	1.958	1.033-3.710
X_{32}: Children	0.546°	0.291	1.726	0.976-3.053	0.849**	0.310	2.336	1.273-4.286
X_{33}: Working other regions/abroad					0.604°	0.310	1.830	0.996-3.362
Constant	2.457***	0.621	0.086		3.503***	0.685		
Nagelkerke pseudo R^2	0.245***				0.248***			

Significance level: *** <0.001; **<0.01; *<0.05; °<0.10

The selected predictors for Model 1 are a subset of those included in Model 2. This shows that severe depression absorbs a large part of the predictive capacity of other predictors. Indeed, there is an opposite direction of the statistical relationships between depression and whole set of predictors selected in Model 2 and between FTP and the same predictors. This led us to analyse in depth the regressors selected in Model 2, instead of Model 1, with the mental caveat that severe depression strongly inhibits youth's FTP.

Severe depression is a main predictor of negative FTP. In fact, high rates of depression have been found among young adults with a severe perception of dark future for themselves. In line with Bandura (1990), we can state that a non-depressed person is more focused on future outcomes, but individual outcomes are also expected to depend on personal abilities. This indicates an optimistic self-concept and a belief in self-efficacy.

The selected predictors in Model 2 can be classified into three main categories: personality traits, aspirations and social capital, with severe depression as an additional, standing-alone predictor. The personality traits showing a relation with a positive FTP are, in order of relevance: self-effectiveness, optimism, and conscientiousness. Self-effectiveness—expressing the self-perceived skill to behave in concrete ways and achieve results—is the personality trait most correlated with youngers' hedonistic future time perspective (Rosenberg, 1985). Being correlated with the propensity to take risks, a higher dose of FTP may pinpoint young people's propensity to start their own business in the future, make financial choices and take investment decisions (Sekścińska et al., 2018; Kiani et al, 2020).

Optimism reflects the self-perceived disposition that more positive than negative things will happen in the future (Luthans et al., 2007; Renaud et al., 2019). Research has shown that optimism is a relatively stable trait in individuals over long periods of time, which may explain its relationship with future prospects. Indirectly, this implies that both optimistic views may promote personal achievements and achievements in turn entail the expectation of predominantly positive events.

Conscientiousness—considered in the Anglo-American literature as a personality trait showing high positive prediction of individual FTP (Zimbardo and Boyd, 1999; Kooij et al., 2018) —is fully subrogated by self-effectiveness and optimism in our analysis and its residual is unexpectedly negatively related to FTP. We could guess that hard-working people who have an ineffective or pessimistic personality are inhibited in their future thinking.

As far as young adults' expectations are concerned, those highly related to FTP are family affection and parenting. Other expectations relate to achieving concern to the obtainment of a degree and traveling, which is typical of the age group at hand. An interesting relation with FTP is expectation to work in outer Italian regions or abroad, which means that people who are open to work experiences far-from-home are much more likely to behave proactively than those whose work ambitions remain circumscribed to the local labour market. All of these expectations are gender-independent, in the sense that gender is statistically significant in the FTP explanation when all of these predictors entered Model 2. The non-significance of gender in Model 1 means that women are much more exposed to mental disorders than men and this not only obfuscates their future orientation but may also inhibit their behaviour.

4. Conclusion

We measured FTP with eight dichotomous items and found, through factor analysis, that our 8-item FTP represents a one-dimensional trait of young adults. Hence, this procedure can be considered as a friendly diagnostic tool to define individual FTPs.

Our study has shown that Italian young adults tend to possess a sound disposition to hedonistic future times. This disposition is higher in people with a self-effective and optimistic personality, in those whose hopes and wishes are higher, and in those who have a richer familial capital, while it is much lower in those with depression and/or a pessimistic personality and/or scanty personal expectations. These findings are in line with the mainstream literature. However, Morselli (2013) stated that recent societal conditions, like the economic crises in the Western world, may change the meaning itself of the future, moving from hope to threat. This explains the relevance of severe depression in explaining the observed youth's FTP variability but may also mean that further research on FTP is needed, particularly as youngsters are concerned.

What is somewhat new is that youngsters' anticipated futures imply a deep concern for family settling. This concern is gender-independent: in fact, it significantly predicts FTP even when gender is selected for the model. The fact that the majority of proactive young adults are prepared for family life may be promising for Italian demographic policies, Italy being a country characterized by a very low birth rate.

Another relevant relation is the higher FTP of men compared to women. This relation is significant only if depression is ignored as an FTP predictor. In other words, there is no relation between gender and future orientation if we leave aside the severely depressed young adults—the vast majority of whom we know to be women—and statistically analyse the other people (this analysis was indeed conducted but its results are not fully reported due to the limited number of pages).

Definitely, there is a hidden variable that may inhibit women's ability to better perceive their futures. The literature suggests that it could be either because women are less inclined to take risks, avoid failure or engage in competitive situations (Croson and Gneezy, 2009), or the conjecture that in order to envision their future, women start from their social role, which is a combination of education, career, family and leisure, whereas men only envision one or two relevant pursuits for their future (Greene and DeBacker, 2004). So, there is still room for future research. In any case, this suggests that social motivation should be enhanced across genders.

Our work may suffer for small size and the convenience recruitment of the sample. This could limit the generalisability of the obtained results. Of course, the repetition the survey in the future could confirm (or disconfirm) the outcomes of our analyses and/or assess the effects of possible social interventions to enhance the FTP of Italian youth. Besides, such large proportion of people showing severe depression as those found in our research poses a relevant problem of youth's mental health. Indeed, social and health authorities should reflect on the fact that depression could be a severe consequence of a difficulty for youngsters to perceive their social and hedonic future.

References

Andre, L., van Vianen, A.E.M., Peetsma, T.T.D., Oort, F.J. (2018). Motivational power of future time

perspective: Meta-analyses in education, work, and health. *PLoS One*, **13**(1): e0190492.
Bandura, A. (1990). Perceived self-efficacy in the exercise of personal agency. *Journal of Applied Sport Psychology*, **2(**2), pp.128-163.
Carstensen L.L., Isaacowitz D.M., Charles S.T. (1999). Taking time seriously. A theory of socioemotional selectivity. *The American Psychologist*, **54**, pp.165–181.
Costa, P.T.Jr, McCrae, R.R. (1978). Age differences in personality structure revisited: Studies in validity, stability, and change. *International Journal of Aging and Human Development*, 8(4), pp. 261-275.
*Croson, R., Gneezy, U. (*2009). Gender differences in preferences. *Journal of Economic Literature,* **47**(2), pp. 448-474.
Di Zio, S., Fabbris, L. (2023) Repression of the future-oriented disposition of Italians by a never-ending pandemic. In: E. Di Bella, L. Fabbris., C. Lagazio (eds.) *ASA 2022 Data-Driven Decision Making*, *Book of Short Papers of the On-site Conference.* Genoa: Genova University Press - Firenze University Press.
Greene, B.A., DeBacker, T.K. (2004). Gender and orientations toward the future: Links to motivation. Educational Psychology Review, 16(2), pp. 91-120.
Hosmer, D.W., Lemeshow, S. (2000). Applied Logistic Regression (II ed.). New York: John Wiley & Sons. IBM (2017) *SPSS Statistical Package for Social Sciences for Windows*, Ver. 25. IBM Corp., Armonk, N.Y., USA.
Kiani, A., Liu, J., Ghani, U., Popelnukha, A. (2020). Impact of future time perspective on entrepreneurial career intention for individual sustainable career development: The roles of learning orientation and entrepreneurial passion. *Sustainability*, **12**(3864).
Kooij, D.T.A.M., Kanfer, R., Betts, M., Rudolph, C.W. (2018) Future time perspective: A systematic review and meta-analysis. *Journal of Applied Psychology*,**103**(8), pp. 867-893.
Lang, F.R., Carstensen, L.L. (2002). Time counts: Future time perspective, goals, and social relationships. *Psychology and Aging*, **17**(1), pp.125-139.
Luthans, F., Avolio, B.J., Avey, J.B., Norman, S.M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology*, **60**(3), pp. 541–572.
Mazzotti, E., Fassone, G., Picardi, A., Sagoni, E., Ramieri, L., Lega, I., Camaioni, D., Abeni, D., Pasquini, P. (2003). II Patient Health Questionnaire (PHQ) per lo screening dei disturbi psichiatrici: Uno studio di validazione nei confronti della intervista clinica strutturata per il DSM-IV asse I (SCID-I). *Italian Journal of Psychopathology, **9***(3), pp. 235-242.
Morselli, D. (2013). The olive tree effect: Future time perspective when the future is uncertain, *Culture & Psychology*, **19**(3), pp. 305-322.
Peetsma, T., Van der Veen, I. (2011). Relations between the development of future time perspective in three life domains, investment in learning, and academic achievement. *Learning and Instruction*, **21**(3), pp. 481-494.
Rosenberg, M. (1985). Self-concept and psychological well-being in adolescence. In R.L. Leahy (ed.). *The Development of the Self*. Orlando, FL: Academic Press, pp 205-246.
Sekścińska, K., Rudzinska-Wojciechowska, J., Maison, D.A. (2018) Future and present hedonistic time perspectives and the propensity to take investment risks: The interplay between induced and chronic time perspectives. *Frontiers in Psychology*, **9**(920).
Spitzer, R.L., Kroenke, K., Williams, J.B. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *Journal of the American Medical Association*, **282**(18), pp. 1737-1744.
Trommsdorff, G. (1994). Future time perspective and control orientation: Social conditions and consequences. In Z. Zaleski (Ed.), *Psychology of Future Orientation* (pp. 39–62). Wydawnictwo Towarzystwa Naukowego Katolickiego Uniwersytetu Lubelskiego.
Zelenski, J.M., Larsen, R.J. (2002). Predicting the future: How affect-related personality traits influence likelihood judgements of future events. *Personality and Social Psychology Bulletin*, **28**(7).
Zimbardo, P.G., Boyd, J.N. (1999). Putting time in perspective: A valid, reliable individual-differences metric. Journal of Personality and Social Psychology, **77**, pp.1271–1288.

doi.org/10.26398/asaproc.00107

Travel and vacations of Italian senior tourists

Angela Maria D'Uggento[1], Francesco Domenico d'Ovidio[1],
Claudia Marin[2], Ilaria Pepe[1], Ernesto Toma[1]
[1] Department of Economics and Finance, Aldo Moro University of Bari, Italy
[2] Department Education, Psychology, Communication Aldo Moro University of Bari, Italy

1. Introduction

The number of older people is growing rapidly. The United Nations (2015) estimates that, by 2050, more than two billion people will be 60 years and older, accounting for 22% of the world's population while it was 10% in 2000, steadily increasing over the last century in almost all industrialized countries worldwide (Harper and Leeson, 2008; Meiners and Seeberger, 2010; Harper, 2014). Aging represents one of the most significant societal changes for all sectors of society, including labor and financial markets, housing, transportation and social protection, as well as in family structure and intergenerational relationships.

Much of the concern about population aging is based on assumptions arising from the characteristics of today's older population, but the scenario could change. Older adults are more likely to remain active, productive and able to contribute to the society because they have greater human capital in the form of education, skills and even a good health profile. Travelling is an important activity that can improve the health profile, ease the transition to retirement and improve the quality of life of seniors. So, the tourism industry is increasingly targeting seniors as an important market segment due to their frequent travelling and significant spending, and the demand for tourism and leisure activities among seniors is expected to grow. These activities are now recognized as important aspects of seniors' lives and contribute to both mental and physical well-being (Silverstein and Parker, 2002; Jang and Wu, 2006).

The literature review in Patterson and Balderas (2020) includes several studies on senior tourism, travel behavior, and motivations published between 1994 and 2018. Some researchers have focused on the finding that the senior tourism market segment is very heterogeneous, making it difficult to treat them as a unified group. Research by Lehto et al. (2002, 2008), Moschis (2003), and Nimrod and Rotem (2012) shows that older travelers have different preferences, behaviors and motivations that distinguish them from younger travelers. Reece (2004) emphasized that seniors differ from younger people in their reasons for travel and their choice of destination. Other studies have shown that older travelers from different generations have different travel inclinations, participation in destination activities and travel philosophies (Szmigan and Carrigan, 2001). Patterson (2018) observes that this heterogeneity implies a wider range of travel options to cater the preferences of seniors. These options range from soft adventure trips, which seniors may prefer to organize themselves via the internet, to fully organized group trips arranged by travel agencies and including stays in five-star hotels.

It is to be emphasized the lack of consensus regarding the definition of seniors in the literature. Möller et al. (2007) divide seniors into three life stages: 1. empty nesters (55-64 years); 2. young seniors (65-79 years); 3. seniors (80 years and older). The first group includes people who are still working but whose children have left home and are no longer financially dependent on them. They are generally less indebted and have sufficient means to finance their needs, so they can afford luxury goods and travel frequently but for shorter periods of time.

The "young seniors" are those who have recently retired and fall into the time rich category. They rely on their past savings for current expenses and are very health conscious. Unless they have serious health problems, they prefer to travel and spend more on high-end goods and services.

People in the late retirement phase are those whose health often declines and have an increased need for medical care or retirement homes. They travel less frequently and prefer domestic destinations.

If this classification stems from marketing theory, a further debate arises about the definition of seniors based on age, as different age limits used by tourism researchers cause confusion: some of them consider people aged 50 as seniors. However, gerontological research is more consistent and usually defines seniors as people aged 65 and older. Other researchers divide the older population into two groups: those aged 50 to 65 and those aged 65 and older, with the former being the most frequent travelers and therefore of greatest interest to the tourism industry. Older travelers have characteristics and preferences that distinguish them from younger travelers. First, generational differences result in different behaviors and preferences, travel philosophies and participation in activities than younger generations. Because seniors have different reasons for traveling, they choose different destinations than younger travelers and their decisions are influenced by factors such as health, mobility and stage of life.

Due to different preferences, seniors need a wide range of travel options. Some prefer adventure travel that they can plan independently, while others opt for organized group trips with all-inclusive packages. Travel and leisure activities play a crucial role in seniors' well-being. Participating in these activities can improve both mental and physical health. In addition, travel can help seniors adjust to the transition to retirement by giving them a sense of purpose and engagement.

The aim of our study is to identify the main characteristics of older tourists, their behavior and their motivations using data from the Viaggi e vacanze survey (Istat, 2023).

2. Data and methods

Viaggi e vacanze is part of the ISTAT household expenditure survey of 2014 and enables the collection of information on the tourist movements of Italian residents. The data refer to the number of tourists, trips, overnight stays on trips and excursions in the country or abroad. From 1997 to 2013, this information was collected as part of a quarterly survey entitled Travel, vacations and daily life.

For statistical purposes, tourism is defined as the activities and services in which people move outside their "usual environment" for vacation or business purposes. Tourism flows, therefore, include all non-habitual trips with overnight stays (travel) or without overnight stays (excursions). The identification of a person's usual environment makes it possible to correctly distinguish the tourism phenomenon from mobility that does not fall within the scope of tourism demand. Habitual trips and excursions, i.e. trips made weekly to the same place that is different from the place where one lives, are nevertheless classified as habitual environment and do not fall under tourist flows; such trips are in fact considered part of the individual's daily life and habits. Also excluded from the definition of "tourist" are people who travel on a daily or weekly basis for work, study or personal reasons, i.e. when travel is part of routine activities. Tourist trips are classified according to international standards by distinguishing between trips for work purposes and vacation trips, and between 'short' (one to three nights) and 'long' (more than three nights) vacation trips. Vacation trips include trips for leisure, pleasure, relaxation, visiting relatives or friends, for medical treatment or for religious reasons.

The survey is based on a random sample and is conducted every month of the year. The sample design is two-stage, with the first stage being stratified: first-stage units are municipalities, second-stage units are households. In 2022, a total of 542 municipalities were involved. The sample design included a theoretical annual sample of approximately 32,500 households living in the 224 municipalities that participated in the survey each month. The survey data for 2023 were collected randomly in 2022 using the national register of residents, following a sampling strategy aimed at obtaining a representative sample of households residing in Italy.

The survey was conducted using a CAPI (Computer Assisted Personal Interview) technique. Each person in the household was asked about the trips and excursions undertaken during the reference period of the survey. The dataset published by ISTAT includes a total of 5,429 respondents, of which

we have selected those aged 65 and over, who make up 16.6% of the sample (901 respondents), according to the gerontological definition. In our analysis we consider two subgroups: travelers aged 65 to 74 (650 respondents, 72.1% of the target population) and travelers aged 75 and older (251 respondents, 27.9%).

The data were analyzed using classification and regression trees (CRT) techniques (Breiman et al. 1984; Hastie et al. 2013) to obtain a segmentation of older travelers based on their main characteristics. The decision process of a classification tree is straightforward as it splits the sample according to a variable at a time and produces a clear, hierarchical structure of splits. This process is repeated recursively at each node. A classification tree determines the threshold for splitting both numeric and categorical variables by evaluating the splits that maximize the difference between the two resulting groups while making each group as homogeneous as possible.

The quality of a split is based on an "impurity index" (e.g. Gini index, entropy) for each group and combines them into a weighted average impurity for the split. The algorithm selects the threshold that minimizes this impurity and ensures that the groups are as homogeneous as possible with respect to the target variable. In this way, it maximizes the separation between the classes in the two groups. Statistical analyses were performed using IBM SPSS software (IBM, 2017).

3. Results and discussion

Our target population is divided into two groups: "young seniors" aged 65-74 (72.1%, 650 respondents) and "old seniors" aged 75 and over (27.9%, 251 respondents). The sample consists of 72.1% women, who are slightly more frequent among the young seniors (74.0%). In terms of marital status, widowed people predominate in the ≥75 age group (26.3% compared to 8.0% in the other age group), while in the 65-74 age group around three quarters of respondents are married (74.9% compared to 62.5% among older tourists), divorced (12.0% compared to 4.0%) or unmarried (5.1% compared to 7.2%). Young seniors have a higher level of education: 27.5% (19.9%) have a college degree and 41.8% (45.4%) have a high school diploma.

It is interesting to examine the main reasons for traveling, which have almost the same proportion in the two subgroups: "vacation, having fun" is the most chosen option (61.6% in the total sample; about 61.9% among young seniors *vs.* 60.7% in the other group), followed by "meeting relatives/friends" (33.6% in the total sample; about 33.7% among young seniors and 33.6% in the other group). Health care, wellness and medically prescribed treatments account for only 4.0% of the sample, but the proportion rises to 4.5% among older travelers (compared to 3.8% in the 65-74 age group).

The decision tree provides a detailed analysis of the characteristics of older travelers by segmenting them according to key variables such as type of travel arrangements, type of accommodation, destination, total expenditure and trip duration (Figure 1). Traditional sociographic variables such as marital status, gender and occupational status were excluded from the set of independent variables for the CART as, although they played a primary role in creating the tree, they were not found to be relevant in profiling the behaviors of older travelers. Education was included as it was considered a proxy for income and cultural level of respondents.

The classification tree in Figure 1 shows how the different variables influence travel decisions with different weightings. The tree has 12 nodes, of which 7 are terminal nodes. The first subdivision is defined by the use of the internet for booking accommodation and transfers: The overall sample is almost evenly distributed, with a slight prevalence of non-users (53.4%, node 2), but the proportions within the two groups remain quite similar to those in node 0.

The second important variable is the main accommodation type: in both paths, a larger proportion of respondents aged 65-74 are found among tourists who mainly opt for house rentals, bed and breakfast and other private accommodation (approx. 85.0%), while the other two nodes (4 and 5) have a larger proportion of tourists aged 75+ who prefer health and care facilities, multi-owner accommodation and accommodation with friends and relatives. While the path of those who book via

the Internet ends with the breakdown by main accommodation, the path of non-users provides additional information on around 400 respondents.

In terms of destination, only a small proportion of the total sample (5.8%) traveled abroad, mainly people in the 65-74 age group (82.7%). Around a third of the sample (36.8%) prefer to travel in Italy, with 37.0% belonging to the older travelers. The group of older tourists who spend their vacation in Italy is further subdivided based on the length of stay. Those taking a longer trip, i.e. more than 2 weeks, make up only 5.1% of the sample, with a high proportion of respondents aged 65-74 (approx. 83.0). For 31.7% of the total sample (node 9), the vacation lasts 2 weeks or less and 40.2% of these are older people over the age of 75.

The final variable used to segment the sample is the average total cost of the vacation. The threshold value of the splitting variable "Average overall expenses" is 1,125.0 euros: most respondents (30.4%, node 11) are below this, including 6 out of 10 respondents who belong to the 65-74 age group. The last node is characterized by higher average total expenditure and, although it is very small in terms of numbers, consists almost exclusively of people aged 75 and over.

Table 1 shows the relative importance of the individual attributes for the classification. The most important attribute "use of the internet for bookings" is assigned a normalized importance of 100, and all other attributes are scaled relative to this value. The second most important attribute is the main type of accommodation, followed by other organizational aspects of the vacation such as the total cost, the duration, the organization of the accommodation (direct booking or via a travel agency), the destination, the average daily cost and the visit to a historic city. The last two variables belong to the individual characteristics such as the level of education and, less importantly, the motivation for the trip.

4. Final remarks

The analysis of data allowed us to unveil the organizational modalities of Italian senior tourists. We found that half of Italian senior tourists use the Internet to get insight and/or book their travel. In general, 3 out of 10 elderly book without internet and spend their vacation in hotels, with friends/relatives, in owner-occupied or multi-family homes or in facilities offering health and care services. They almost always stay in Italy for a period of up to two weeks and usually spend just over 1,000 euros.

It is interesting to see how certain aspects related to the organization of the vacation (type of accommodation, cost and duration of the trip) impact on the travel decisions of older people. This information is crucial for tour operators as they can use this data to tailor their services and offers to the specific needs of the older tourist market segment.

The increasing recognition of the importance of travel and leisure for seniors has significant implications for the tourism industry. Tourism providers need to understand and cater to the different needs and preferences of older travelers, the changing socio-demographic characteristics of the senior market and the benefits of leisure travel. This may mean offering a wide range of travel options, from independent travel to fully organized group tours. In this way, the tourism industry can enhance seniors' travel experiences and contribute to their overall wellbeing and quality of life.

Although we have adopted the gerontological definition, we believe that the most important concept is subjective or "perceived" age (Barak and Schiffman, 1981), which is based on how people perceive their age. Many seniors feel younger than their chronological age (Armenta et al., 2017). This subjective age may be a more reliable predictor of consumer behavior and physical health than chronological age (Stephan et al., 2015).

A limitation of this study arises from the data collected in the ISTAT survey, which relates to organizational aspects of travel, but not to motivations and travel experiences. A questionnaire that also includes such "emotional" experiences and benefits of travel could close this gap.

Figure 1. Classification tree

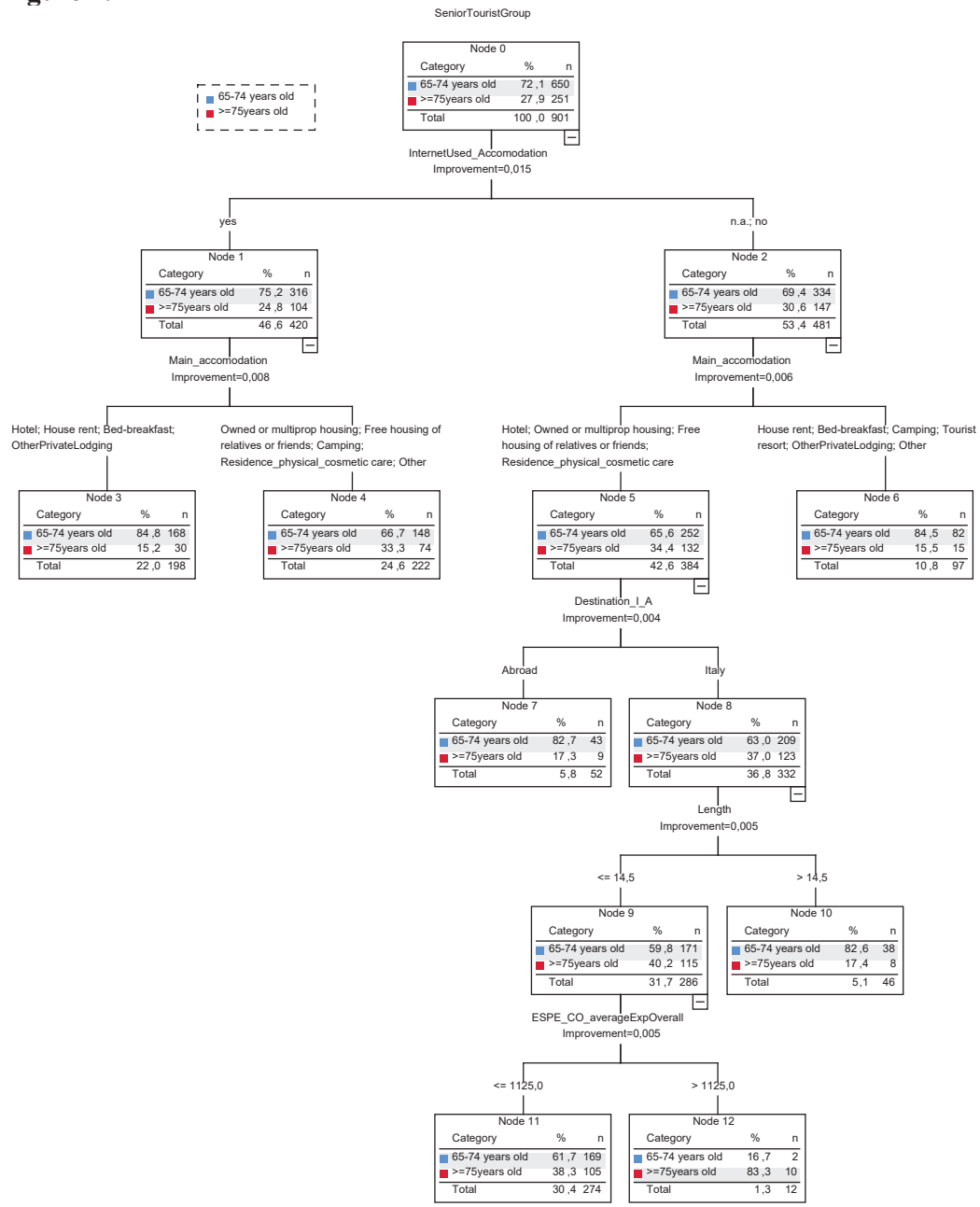


Table 1. Measures of variable importance (%)

Independent Variable	Importance	Normalized importance (%)
Internet for organization/accommodation	0.016	100.0
Main type of accommodation	0.015	92.1
Total cost	0.011	70.4
Length of the trip	0.008	47.9
LodgingOrganization	0.006	37.6
Destination Italy or abroad	0.006	34.8
Lodging organization	0.003	14.7
Average daily expenses	0.004	27.8
City	0.004	22.9
Education	0.002	15.6
Main reason for travelling	0.002	14.5

Acknowledgements: This study is part of the Progetto di Ricerca di Rilevante Interesse Nazionale (PRIN) P2022B3NFH- The future of sustainability: CUP H53D23009470001.

References

Armenta, B.M., Stroebe, K., Scheibe, S, Postmes, T., Van Yperen N.W. (2017). Feeling younger and identifying with older adults: Testing two routes to maintaining well-being in the face of age discrimination. *PLoS ONE* **12**(11): e0187805.

Barak, B., Schiffman, L.G. (1981). Cognitive age: A nonchronological age variable. *ACR North American Advances.*

Breiman, L., Friedman. J.H., Olshen. R.A., Stone. C.J. (1984). *Classification and Regression Trees.* Wadsworth. Belmont. CA.

Harper, S. (2014). Economic and social implications of aging societies. *Science,* **346**(6209), pp. 587–591.

Harper, S., Leeson, G. (2008). Introducing the journal of population ageing. *Journal of Population Ageing*, **1**(1), pp.1–5.

Hastie, T., Tibshirani, R., Friedman. J. (2013). *The Elements of Statistical Learning*. Springer series in Statistics (1st ed.). Berlin: Springer.

IBM (2017) *SPSS Statistical Package for Social Sciences for Windows, Version 25*. IBM Corp., Armonk, N.Y., USA.

Istat (2023). *Viaggi e vacanze in Italia e all'estero*. www.istat.it

Jang, S., Wu, C.M. (2006). Seniors' travel motivation and the influential factors: An examination of Taiwanese seniors. *Tourism Management*, **27**, pp. 306–316.

Lehto, X.Y., O'Leary, J.T., Lee, G. (2002). Mature international travelers: An examination of gender and benefits. *Journal of Hospitality & Leisure Marketing,* **9**(1–2), pp. 53–72.

Lehto, X.Y., Jang, S.S., Achana, F.T., O'Leary, J.T. (2008). Exploring tourism experience sought: A cohort comparison of baby boomers and the silent generation. *Journal of Vacation Marketing*, **3**, pp. 237–252.

Meiners, N., Seeberger, B. (2010). Marketing to senior citizens: Challenges and opportunities. *The Journal of Social, Political, and Economic Studies,* **35**(3), pp. 293–328.

Möller, C., Weiermair, K., Wintersberger, E. (2007). The changing travel behaviour of Austria's ageing population and its impact on tourism. *Tourism Review*, **62**(3/4), pp. 15–20.

Moschis, G.P. (2003). Marketing to older adults: An updated overview of present knowledge and practice. *Journal of Consumer Marketing,* **20**(6), pp. 516–525.

Nimrod, G., Rotem, A. (2012). An examination of the innovation theory of successful ageing among older tourists. *Ageing & Society*, **32**(3), pp. 379–404.

Patterson, I. (2018). *Tourism and Leisure Behaviour in an Ageing World*. Oxfordshire: CABI.

Patterson, I., Balderas. A. (2020). Continuing and emerging trends of senior tourism: A review of the literature. *Population Ageing,* **13**, pp. 385–399.

Reece, W. (2004). Are senior leisure travelers different? *Journal of Travel Research*, **43**, pp. 11–18.

Silverstein. M., Parker, M.G. (2002). Leisure activities and quality of life among the oldest old. *Sweden Research on Aging*, **24**(5), pp. 528–547.

Stephan, Y., Sutin, A.R., Terracciano, A. (2015). How old do you feel? The role of age discrimination and biological aging in subjective age. *PLoS One*. **4,**10(3): e0119293.

Szmigan I., Carrigan, M. (2001). Learning to love the older consumer. *Journal of Consumer Behaviour*, **1**(1), pp. 22–34.

United Nations. Department of Economic and Social Affairs. Population Division. (2015). *World Population Prospects*. New York: USA.

doi.org/10.26398/asaproc.00108

It's how you use the items that counts: An intelligent procedure for item selection in item response theory

Ottavia M. Epifania[a,b], Pasquale Anselmi[b], Egidio Robusto[b]
[a] Department of Psychology and Cognitive Science, University of Trento, Italy
[b] Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Italy

1. Introduction

Item response theory (IRT) models allow for the estimation of the probability that an individual will endorse an item, based on their latent trait level and the characteristics of the items, as determined by various parameters. The details at the item level provide insights into the precision with which each item measures different levels of the latent trait. Consequently, IRT models are particularly valuable for developing short test forms (STFs), as they allow for minimizing the number of items administered while maximizing measurement precision.

Various procedures exist for developing STFs based on IRT models. These procedures aim either at the adaptive selection of items (i.e., each person is administered a different subset of items tailored to their specific latent trait level) or at the fixed selection of items (i.e., static STFs, where all individuals receive the same subset of items). This contribution focuses on the development of static STFs. Therefore, unless otherwise specified, the term "STFs" refers to static STFs.

Recently, Epifania, Anselmi, and Robusto (2022) introduced a new procedure for the development of STFs. This procedure is based on the a priori definition of the so-called θ targets, which are the levels of the latent trait on which the measurement precision should be focused the most. Although the procedure proved its soundness in developing STFs able to precisely measure the entire latent trait, the number of items to include in the STFs must be known in advance (it actually equals the number of θ targets).

In this contribution, a new procedure is introduced, which aims at developing STFs with characteristics that match those of a theoretically defined information function describing the characteristics of the desired test. As such, the number of items does not need to be set in advance. The procedure is validated through simulation studies and its performance is compared against a "brute force" procedure.

The paper is organized as follows: Section 2 presents the 2-Parameter logistic model (2-PL, Birnbaum, 1968) and the information functions of the items and of the test. The descriptions of the newly introduced procedure and of the brute force procedure follow. In Section 4 a simulation study is presented and in Section 5 its results. A brief section illustrating the potentials and pitfalls of this procedure concludes the argumentation.

2. Item and test information function in item response theory

Different IRT models are available according to the nature of the observed responses (i.e., dichotomous vs. polytomous) and to the level of granularity desired for describing the functioning of the items (i.e., the number of item parameters). In this application, we refer to the 2-PL model (Birnbaum, 1968) for dichotomous responses.

According to the 2-PL model, the probability of observing a correct response on item *i (i = 1, 2, . . .N)* provided by person *p (p = 1, 2, ... , P)* is:

$$P(x_{\{pi\}} = 1 | \theta_p, b_i, a_i) = \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}$$

where θ_p describes the latent trait level of person p (also known as the *ability* parameter), b_i is the *difficulty* of item i (i.e., the location of the item on the latent trait), and a_i is the *discrimination* parameter of item i, which describes the ability of the item to discriminate between respondents with different θ levels.

The probability of a correct response depends on the distance on the latent trait between the person parameter θ_p and the item location b_i, such that if $\theta_p > b_i$, this probability is greater than 0.50, while if $\theta_p < b_i$, the same probability is lower than 0.50. This probability is further influenced by the discrimination parameter, such that the higher the value of a_i, the more two respondents with similar θ levels have different probabilities of endorsing the item.

The *item information function* (*IIF*),

$$IIF_i = a^2_i[P(\theta, b_i, a_i)(1 - P(\theta, b_i, a_i))],$$

expresses the precision with which each item measures each level of the latent trait θ. In the 2-PL model, the IIF is strongly influenced by the discrimination parameter (i.e., the higher the value a_i of item i, the higher its *IIF*), and it reaches its maximum (i.e., the item is mostly precise) when the ability θ_p matches the location b_i of item i, such that $P(\theta, b_i, a_i) = 1 - P(\theta, b_i, a_i)$.

By summing up the *IIF* of all the items, the *test information function* (*TIF*), $TIF = \Sigma_i IIF_i$, is obtained, which expresses the measurement precision of the test as a whole. Being the sum of *IIF*s, the *TIF* strongly depends on the number of items included in a test. It can be increased just by adding new items. Given that the aim of the procedure introduced in this study is to minimize the number of administered items while recovering as best as possible the characteristics of a specific test, the use of the *TIF* as a criterion for choosing the most informative STF is misleading and would risk favoring STFs with a higher number of items.

As such, the mean TIF,

$$\overline{TIF} = \sum_i IIF_i / N,$$

is used as a criterion for selecting the most appropriate STF. The logic behind this is that the more items included in the test, the greater the penalization of the TIF. In other words, for each θ level, and assuming the same IIFs, STFs with fewer items would be preferred.

3. Item selection procedures

In what follows we present two item selection procedures: (i) the *item locating algorithm* (ILA), which is the procedure aimed at recovering a specific TIF determined a priori by locating the most appropriate items from the item bank, and (ii) the *brute force procedure* (BFP), which selects the STF that best recovers the TIF determined a priori after trying all the possible combinations of items of different lengths. Both ILA and BFP require the definition of a TIF target, which describes the desired characteristics of the STF, regardless of its length. Both procedures attempt at recovering as best as they can the TIF target, denoted as *TIF**.

The BFP was introduced to check whether ILA is able to recover the best possible combination of items among all the possible combinations of items of different lengths that can be obtained from the item bank. The simulation study compares the performance of ILA and that of BFP in 100 simulations. The use of a brute force algorithm, like BFP, allow for obtaining the best possible item combination among all the possible item combinations. As such, the item selection provided by BPF is considered as the gold standard against which the results of the other algorithm can be compared. Further details are provided in Section 4.

3.1 Item locating algorithm

ILA is an iterative algorithm that compares the *TIF** with temporary *TIF*s, denoted as TIF_{TE}, obtained at each iteration by adding one item at a time. ILA iterates until the absolute difference between *TIF** and the TIF_{TE} that includes the last selected item is equal to or greater than the difference between *TIF** and the TIF_{TE} that does not include the last selected item (i.e., termination criterion).

The TIF_{TE} is the mean temporary *TIF* obtained as:

$$TIF_{\{TE\}} = \frac{\sum_{\{i \in Q\}} IIF_i}{||Q||}$$

where *//Q//|* denotes the cardinality of set *Q* that contains the indexes of the items selected for inclusion in the STF up to that iteration.

The algorithm iterates the following steps until the termination criterion is reached:

1. The absolute difference between *TIF** and TIF_{TE} is computed since at the beginning no items are selected for inclusion in the STF, $Q=\emptyset$ and $//Q//=0$, so the difference between TIF^* and $TIF_{TE\ is\ 0}$
2. A theta target is determined as the θ level for which the absolute distance between *TIF** and TIF_{TE} is maximum
3. The index of the first item to be included in Q is the index of the item whose location is closest to the θtarget:
4. The mean temporary *TIF* is computed
5. Repeat from Step 1 until the termination criterion is reached.

3.2 The brute force procedure

BFP develops the STFs considering all the possible combinations of items without repetition and compares the STFs from those of length 1 (i.e., those composed of one item only) to those of length *L*, where $L = N - 1$ and N is the number of items included in the item bank.

For each of the possible lengths of the STF, a combination of items is developed. For instance, if the item bank is composed of 10 items, $L = 9$ and $\binom{10}{1} + \binom{10}{2} + \binom{10}{3} + \binom{10}{4} + \binom{10}{5} + \binom{10}{6} + \binom{10}{7} + \binom{10}{8} + \binom{10}{9} = 1{,}022$ will be developed and tested by the BFP.

For each of the possible combinations of items of different length, the TIF is computed, and the absolute difference from the *TIF* target and a mean of the absolute difference are obtained. The best STF is the one with the lowest mean absolute distance from the TIF^*, that is, the one that presents the lowest absolute distance from the TIF target.

4. Simulation study

The simulation study was run in R (R Core Team, 2022). We iterated the following steps for 100 times:

1. Generate an item bank composed of $N = 6$ items with difficulty parameters drawn from a uniform distribution $U(-3,3)$ and discrimination parameters drawn from a uniform distribution $U(0.90,2.0)$;
2. Generate a mean *TIF** by randomly selecting items from the item bank. The number of selected items varied between 2 and 5 (Mean = 3.34 ± 1.13). The starting difficulty and discrimination parameters of the selected items are modified by adding or subtracting values randomly extracted, at each iteration, from uniform distributions $U(-0.20,0.20)$. A constraint is made on the

discrimination parameters such that they cannot result in negative values after they are modified through the addition or subtraction of the value;

3. Considering the TIF defined at Step 2 and the item parameters generated at Step 1:
    - **ILA** searches for the best possible STF able to recover the TIF target.
    - **BFP** looks for the best possible item combination. Given that the item bank is composed of $N = 6$ items, $L=5$, and 62 STFs are developed and compared.

To evaluate the performance of ILA, the following criteria have been followed:

- **Symmetric distance** between the indexes of the items selected by BFP and ILA. Based on the symmetric distance, it is possible to compute:
    - **Sensitivity**: the number of items that are identified by BFP and selected by ILA;
    - **Specificity**: the number of items that have been excluded by both BFP and ILA;
    - **Accuracy**: the sum of specificity and sensitivity divided by the length of the item bank.
- **Rank** of the STF obtained with ILA in the distribution of STFs obtained with BFP, ordered according to their increasing distance from the average TIF of the STFs that best resemble the TIF target.

Moreover, the occurrences in which the cardinalities of Q were equal according to both ILA and BFP ($//Q_{ILA}// = //Q_{BFP}//$|) were counted in each iteration. If $//Q_{ILA}// = //Q_{BFP}//$, then we checked whether the two strategies resulted in the same selection of items. Otherwise, the difference between the cardinalities was computed as $//Q_{ILA}// - //Q_{BFP}//$ and it was averaged across the iterations.

5. Results

In 100 simulations, ILA and BFP resulted in STFs of the same lengths in 57 cases, out of which 41 cases resulted in the same selection of items. In 34% of cases, ILA selected fewer items than BFP, while in 9% of cases, it selected more items than BFP. The weighted mean of the item difference between ILA and BFP was 0.77.

Averaging across the 100 iterations, ILA included the same items as BFP (sensitivity) and excluded the same items as BFP (specificity) in 72% and 85% of the occurrences, respectively. Concerning the percentile rank of the distance from the best STFs identified by BFP, in 74% of instances, ILA produced STFs within the 10th percentile of distance, while the median percentile rank was 3.17. This means that in 50% of the cases, ILA developed an STF that was within the fourth percentile of distance from the best possible STF.

6. Final remarks

This manuscript introduces a new algorithm, the item locating algorithm, designed to facilitate the development of short test forms within an item response theory framework. While ILA holds promise as an effective tool for test shortening, there are several important considerations that must be addressed in future research to fully evaluate its potential.

As it currently stands, ILA selects items based solely on their location in relation to a predefined θ target, primarily focusing on the difficulty parameter of each item. In other words, ILA uses the distance of each item from the θ target to determine its inclusion in the STF, effectively prioritizing items that are most aligned with the targeted latent trait level.

However, this approach overlooks another crucial item parameter: item discrimination. As it stands, when two items are equidistant from the θ target, ILA selects one of them at random

without considering their discrimination parameters or how informative they are for measuring the target θ. This can be problematic, as an item with higher discrimination could offer more precise information about an individual's latent trait level, making it more useful for the test despite being equally distant from the θ target.

To enhance the algorithm's effectiveness, future iterations should incorporate item informativeness into the selection process. By considering both the item location (difficulty) and its discrimination parameter, the algorithm could ensure that each item included in the STF not only aligns with the θ target but also provides a more accurate and reliable measurement of the latent trait. This would help optimize the measurement precision of the STF and improve its overall validity.

Another issue concerns the size of the item bank, which included only six items. This limited number of items poses significant risks to the replicability of the results, as a small item bank can lead to reduced variability and less generalizability of the findings. The modest size of the item bank was primarily a result of the computational burden associated with the BFP, which requires evaluating all possible combinations of items to select the most appropriate STFs. Given the computational challenges, the current study was constrained to a small set of items.

However, this limitation also raises concerns about the robustness of the results and their applicability to broader contexts. To address this issue, future research should explore the use of larger item banks, which would provide more diverse and representative samples of items.

It is also important to note that the BFP was never designed as a tool for test reduction but rather as a benchmark for assessing the performance of ILA. While BFP provides a useful gold standard, the results obtained with ILA in this study should be interpreted with caution, given the limitations posed by the small item bank. Future research involving larger item banks is crucial to fully understand the applicability, precision, and replicability of ILA.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing.

Epifania, O.M., Anselmi, P., Robusto, E. (2022). Pauci sed boni: An item response theory approach for shortening tests. In *The Annual Meeting of the Psychometric Society* (pp. 75–83). https://doi.org/10.1007/978-3-031-27781-8_7

R Core Team (2022). *R: A Language and Environment for Statistical Computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

doi.org/10.26398/asaproc.00109

School socioeconomic status and students' achievements: Exploring territorial heterogeneity through a multilevel regression discontinuity approach

Rosa Fabbricatore[a], Pasquale Sannino[b], Cristina Davino[a]
[a] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy.
[b] Department of Economics and Law, University of Macerata, Macerata, Italy.

1. Introduction

The debate on how socioeconomic status affects students' performance and educational opportunities dates to the 1960s, a period when this issue was first brought to light by what we can consider the pioneering work, the Coleman Report (Coleman et al., 1966). Among the findings of the report, one of the most important is that a student's level of education is correlated with both their family background and the background of other students in the school. Another line of literature has shown that, regardless of the family's socioeconomic status, a significant factor affecting students' academic performance is the average socioeconomic status of the school, highlighting that schools with a higher average socioeconomic status tend to have higher test scores (Rumberger & Palardy, 2005; OECD, 2004). Thus, in a world increasingly leaning towards equality, it is in the interest of institutions to monitor this trend between socioeconomic status and school performance to ensure equal opportunities for students regardless of their socioeconomic conditions, both family and school, aiming to improve education. Indeed, higher education can enable a better standard of living by providing broader access to various job opportunities. Individuals with a stronger cultural background can more easily access to qualified and better-paid jobs, enjoying concrete prospects for professional advancement. On a larger scale, this phenomenon contributes to reducing the income gap between different segments of the population (Abdullah et al., 2015).

In this vein, assessing students' achievements has become of paramount importance worldwide to highlight any disparities between schools, regions, and geographical areas. In Italy, this work is carried out by the INVALSI (National Institute for the Evaluation of the Educational System of Education and Training) through the annual administration of the so-called "INVALSI Tests," which are periodic and systematic assessments of students' skills and knowledge at the end of the second and fifth grades of elementary school, the third grade of middle school, and the second and fifth years of high school. The tests are prepared and provided by INVALSI and are intended by the Ministry of Education to assess students' proficiency in Italian, Mathematics, and English. The purpose of the INVALSI tests is multiple: to verify the level of student learning, monitor school performance, and analyse if there are differences between different schools, regions, or geographical areas of the country.

The present work aims to investigate the impact of the socioeconomic status of the schools attended by Italian students on the mathematics scores obtained in the INVALSI tests, with a particular focus on regional differences. Indeed, Italy is a country characterized by significant regional differences that manifest in many aspects, from the economic aspect to the quality of life up to education. These disparities are particularly evident when comparing the northern regions with those of the south. The study proposes a multilevel regression discontinuity based on the integration of Regression Discontinuity (RD) and Multilevel regression (Hox et al., 2017). On the one hand, the RD design represents a non-experimental strategy for analysing causal effects, which allows us to estimate the causal effect of attending a school with a low socioeconomic status on the results of the INVALSI Mathematics tests. On the other hand, the multilevel specification provides

new insights into Italian territorial heterogeneity by examining regional variations in the effect of school socioeconomic status.

The remainder of the paper is organized as follows: Section 2 focuses on the methodology employed, specifically detailing how the multilevel approach was integrated with the regression discontinuity technique to achieve the research aims. The data used in the analysis are described in detail in Section 3, while Sections 4 and 5 present the empirical results and some conclusions, respectively.

2. Methods

The RD design, introduced by Thistlethwaite and Campbell in 1960, is a highly credible non-experimental technique for analysing causal effects (Cattaneo et al., 2019; Cattaneo et al., 2024; Imbens & Lemieux, 2008). Typically, experimental designs, which are studies aiming to evaluate the effectiveness of a treatment, require a treatment group and a control group that differ only in their treatment status. RD is considered a quasi-experiment because receiving the treatment depends wholly or partly on whether an observable variable (called running variable, forcing variable, or index) exceeds a certain threshold value (called cutoff). Thus, in RD, there are three fundamental components: *i)* X_i, which is the variable on which the assignment mechanism to the treatment or control group is based (assuming that there are n units indexed with $i = 1,2, \dots, n$), *ii)* the cutoff c, which is an externally determined threshold beyond which units are assigned to the treatment group, *iii)* the treatment D_i, which is a binary variable equal to 1 if the unit i is part of the treatment group and 0 otherwise (the treatment can consist, for example, of implementing an economic policy, an intervention in education, etc.). Each unit is assigned a specific value of the running variable, and treatment status is guaranteed to all units whose running variable value exceeds the known cutoff point and not assigned to those whose running variable value does not.

Several approaches can be employed to estimate the treatment effect formally. When the research aim involves the detection of cross-site heterogeneity of the treatment, random effects models (Hox et al., 2017) proved to be valuable. Accordingly, let Y_{ij} be the potential outcome of the unit i belonging to the site j, the two-level regression model with random effects for the intercept and slope can be expressed as:

$$Y_{ij} = \beta_{0j} + \beta_1\left(X_{ij} - c\right) + \beta_{2j}D_{ij} + r_{ij}; \quad (1)$$

Where:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}.$$

In the above model, γ_{00} represents the mean outcome score, across sites, of units with a running variable equal to the cutoff; γ_{20} defines the treatment effect across sites (the parameter we are interested in); u_{0j} and u_{2j} reflect the variation of γ_{00} and γ_{20} between sites, respectively. Finally, β_1 indicates the general effect of the running variable on the outcome.

Notably, in the approach herein proposed, we tried to improve further what is typically done in multilevel RD implementation in education (e.g., Luyten, 2006; Steinmann & Olsen, 2022) by addressing two main issues of the RD method regarding bandwidth selection and unit weighting. Indeed, the underlying idea of RD is that there exists a certain window around the cutoff point in which belonging to the treatment or control group is almost random, which resembles experimental designs. Moreover, considering a certain weighting function ensures that units closer to the cutoff have a higher weight during model estimation. Specifically, our step-wise approach can be summarized as follows: Step 1) estimating a RD model to select the optimal bandwidth; Step 2) assigning a weight to each observation within the bandwidth according to a kernel function; Step

3) estimating a multilevel model as specified in Eq. (1) considering the sub-sample of units within the selected bandwidth and the corresponding weights. The R statistical software can be used for data analysis, exploiting the packages `rdrobust` (Calonico et al., 2014) and `lme4` (Bates et al., 2015) for steps 1 and 3, respectively.

3. Case study on INVALSI Mathematics tests

The study considered the data from the INVALSI Tests[1] ("Prove INVALSI") administered to 13th-grade students, corresponding to the fifth year of high school, between March and May of the 2022/2023 academic year. After a careful preprocessing phase, we have data on 424,400 students. The outcome variable is represented by the score on the INVALSI Mathematics tests, measured using a Rasch model to account for the difficulty level of the questions and to make the performance of students, classes, and schools comparable. The running variable is the socioeconomic status of the school attended by each student, herein defined according to the ESCS (Economic, Social and Cultural Status) index. The ESCS is a standardized index derived from a principal component analysis of three indicators proposed by INVALSI (Campodifiori et al., 2010): *i)* HISEI – the parents' occupational status, *ii)* PARED – the parents' level of education, *iii)* HOMEPOS – the possession of certain material goods. Thanks to the school offices, the ESCS is measured at the individual level and then aggregated into the class and school ESCS by taking the average. The cutoff is represented by the value set by the Ministry of Education and Merit in Ministerial Decree No. 90 of May 19, 2023[2], which is -0.31243 for secondary schools (from this value, the ESCS is considered low). The treatment group consists of all students enrolled in a school with low ESCS, and consequently, the treatment corresponds to attending a school with low ESCS. As mentioned earlier, treatment status is guaranteed for all units whose running variable value exceeds the known cutoff point. In RD designs, it is customary for the treatment group to be on the right and the control group on the left of the cutoff. To achieve this, we multiplied the running variable by -1, ensuring that all students enrolled in a school with low ESCS are on the right side of the cutoff.

In Figure 1, a set of violin plots represents the school ESCS distribution according to the Italian regions that are placed in ascending order based on the median ESCS value. The trend is quite clear: the southern regions are almost all positioned at the lower part of the graph, indicating a well-known geographical disparity.

4. Results

Before implementing the RD, we ensured that our study was suitable by verifying the assumptions. Specifically, we checked the "no sorting around the cutoff" assumption (McCrary, 2008), which states that individuals cannot sort themselves around the cutoff; they cannot decide whether to be to the right or the left of it. Since the running variable is discrete, we tested this assumption using the test proposed by Frandsen (2017) and found no evidence of possible sorting. To select the optimal bandwidth in Step 1, we used the algorithm that minimizes the Mean-squared error (MSE), which accounts for the so-called "bias-variance" tradeoff. In Step 2, observations within the bandwidth were weighted with a triangular kernel function to ensure that observations closer to the cutoff had a greater weight in the analysis. The results obtained from the RD implementation show that the selected bandwidth is 0.175 to the right and 0.175 to the left of the cutoff, reducing the analysis to an effective number of observations of 54,470 on the left and 42,205 on the right.

[1] https://www.invalsiopen.it/

[2] https://www.miur.gov.it/documents/20182/7414469/m_pi.AOOGABMI.Registro+Decreti%28R%29.0000090.19-05-2023.pdf/44e239ff-2151-8c6a-3d45-851a38834ded?version=1.0&t=1690386661238

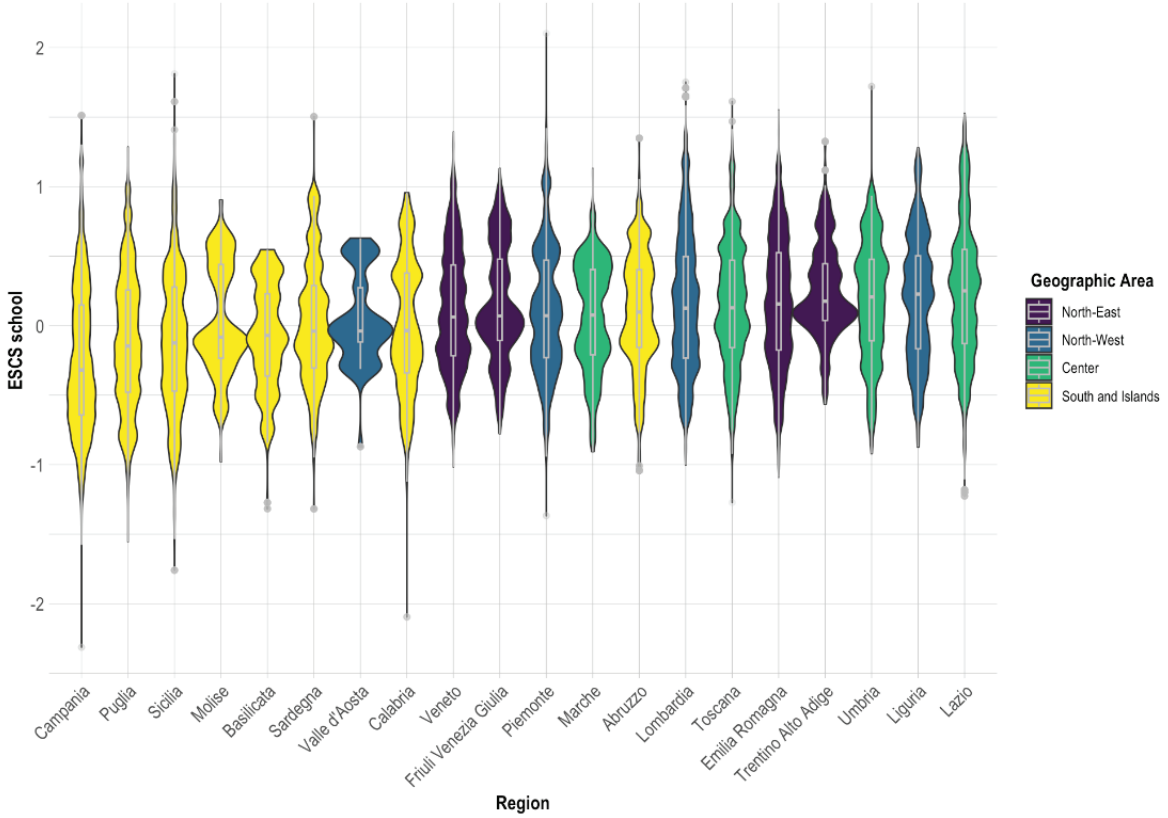


Figure 1 – Violin plots for school ESCS distribution according to Italian regions. Colors indicate the geographical macro-area.

Then, we estimated the multilevel model in Step 3. First, we compared the fully random model specified in Eq (1) (Mod0) with simpler versions, which are the random intercepts model (Mod1) and the null random intercepts model (Mod2). Results reported in Table 1 suggest that Mod0 is preferred for our data, indicating that both intercept and slope (the treatment) significantly vary across regions. Specifically, in Mod0, the fixed effects are equal to $\gamma_{00} = 181.97$ for the intercept, $\beta_1 = -52.29$ for the centered running variable $(X_{ij} - c)$ and $\gamma_{20} = -1.84$ for the treatment effect. Hence, overall, students enrolled in a school with a low ESCS experience a negative causal effect, resulting in a loss of 1.84 points in the INVALSI Mathematics tests. Concerning random effects, the variances of level 2 (regions) residuals are equal to $\sigma^2_{u_0} = 87.11$ and $\sigma^2_{u_2} = 22.19$, with a negative correlation $\rho = -0.65$. Finally, the variance of level 1 (individuals) residuals is equal to $\sigma^2_r = 453.58$.

Table 1 – Fit statistics for multilevel models

Model	#par	AIC	BIC	LogLik	χ^2	p-value
Mod2	3	963,354	963,382	-481,674		
Mod1	5	961,751	961,798	-480,870	1607.00	< 0.001
Mod0	7	961,309	961,376	-480,648	445.49	< 0.001

The caterpillar plots in Figure 2 provide more details on regional differences, showing the estimates of second-level random effects and the related 95% confidence interval. Results reveal that northern regions generally report significantly better performance in the INVALSI Mathematics tests than the overall mean, whereas southern regions present the opposite trend. Moreover, given the negative sign of the treatment effect γ_{20}, the residuals u_{2j} at Level 2 highlight

that being enrolled in a school with a low ESCS has a stronger negative effect on students' performance in the northern regions than in the southern ones. An exception is the Veneto region, which reports a positive residual unlike the other northern regions.

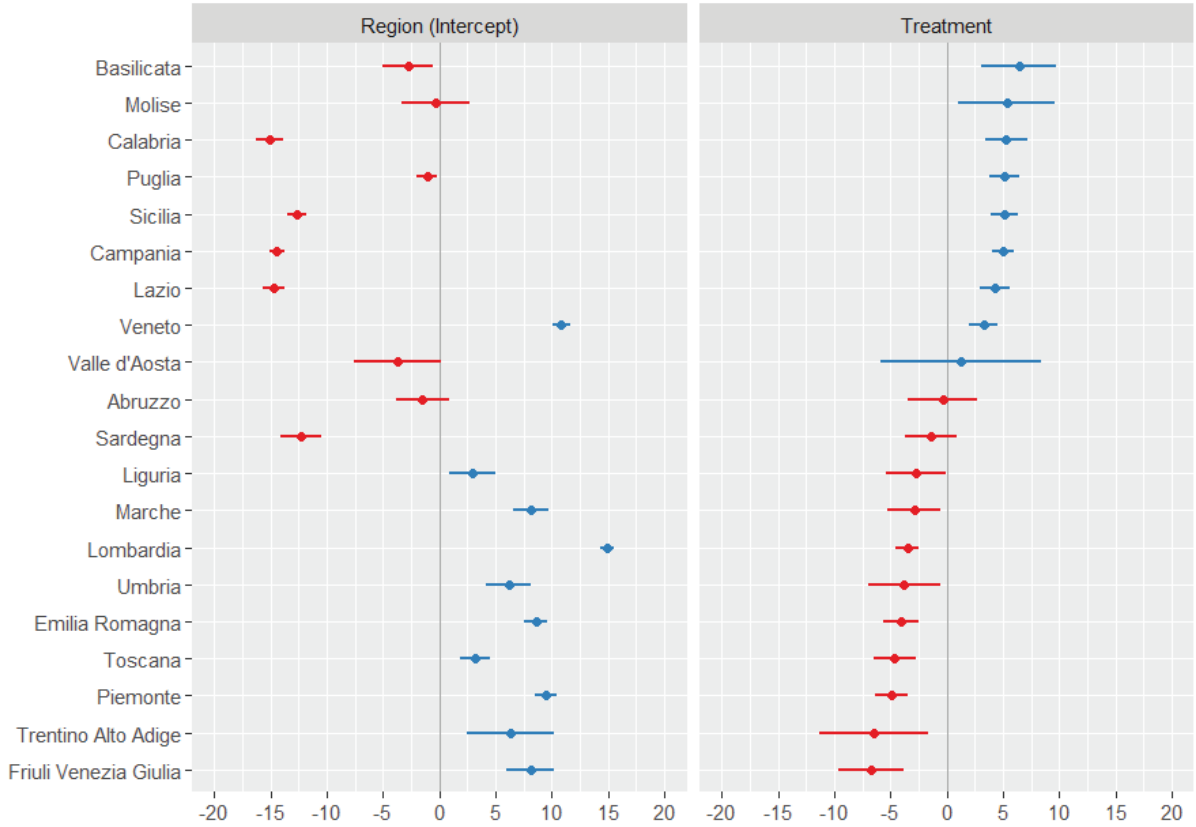


Figure 2 – Caterpillar plots: Estimated second-level random effects with 95% confidence interval for the intercept and the slope (treatment).

5. Conclusion

By integrating multilevel modelling with RD design, this study provides new insights into the interplay between individual, school, and regional factors affecting students' performance in INVALSI Mathematics tests. Results highlight that attending a school with a low ESCS has a negative causal effect on students' performance in mathematics. Moreover, evidence from the territorial heterogeneity in Italy shows that the southern regions generally report the worst mean score in the INVALSI Mathematics tests whereas the northern regions present a stronger negative causal effect of school ESCS on students' achievements. The cause of this more pronounced negative causal effect for regions in the North could be explained by: *i)* a resilience factor: students in the South might have a more resilient outlook towards challenges and disadvantaged contexts; *ii)* a contextual factor: in the South, schools may already operate under generally more unfavourable conditions, and therefore the additional disadvantage associated with a low ESCS index might be less evident, as it is compounded with other pre-existing factors.

This analysis provides significant insights for policymakers, suggesting that a tailored approach to regional needs could be the solution to addressing educational inequalities. Indeed, the emerging regional variations may be influenced by factors such as economic development and educational policies. Future research could focus on either methodological advancement in integrating the three proposed steps into a single-step approach or a more in-depth investigation of the identified regional differences, particularly the uncommon behaviour observed for the Veneto region.

Acknowledgment

Rosa Fabbricatore acknowledges the financial support provided by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP E63C22002140007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

This work was also supported by the University Research Funding Program (FRA) 2022 of the University of Naples Federico II, with the contribution of the Compagnia San Paolo (Title of the project: "Measuring and Mapping Educational Poverty"; principal investigator: Cristina Davino).

References

Abdullah, A., Doucouliagos, H., Manning, E. (2015). Does education reduce income inequality? A meta-regression analysis. *Journal of Economic Surveys*, **29**(2), pp. 301-316.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), pp. 1-48.

Campodifiori, E., Figura, E., Papini, M., Ricci, R. (2010). Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia. *Working paper INVALSI n.2*. Available from: https://www.invalsi.it/download/wp/wp02_Ricci.pdf

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., York, R.L. (1966). *Equality of Educational Opportunity*. Government Printing Office, Washington, DC, (U.S.).

Calonico, S., Cattaneo, M.D., Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, **14**(4), pp. 909-946.

Cattaneo, M.D., Idrobo, N., Titiunik, R. (2019). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press, Cambridge, (UK).

Frandsen, B.R. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. *Regression Discontinuity Designs: Theory and Applications*, **78,** pp. 281-315.

Hox, J., Moerbeek, M., Van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, (U.S.).

Imbens, G.W., Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* **142**(2), pp. 615-635.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education*, **32**(3), pp. 397-429.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142**(2), pp. 698-714.

OECD (2004). *Education at a Glance 2004: OECD Indicators*, OECD Publishing, Paris.

Perry, L.B., McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, **112**(4), pp. 1137-1162.

Rumberger, R.W., Palardy, G.J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record*, **107**(9), pp. 1999-2045.

Steinmann, I., Olsen, R.V. (2022). Equal opportunities for all? Analyzing within-country variation in school effectiveness. *Large-Scale Assessments in Education*, **10**(1), pp. 2.

Thistlethwaite, D.L., Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, **51**(6), pp. 309-317.

doi.org/10.26398/asaproc.00110

Capacity to aspire in high school students: Theoretical definition and measurement

Rosa Fabbricatore, Rosaria Romano, Cristina Davino
Department of Economics and Statistics, University of Naples Federico II, Naples, Italy.

1. Introduction

The high school period represents the time when young people shape their career aspirations, choosing from various educational or working pathways. This process involves the development of a cultural ability known as the "capacity to aspire" (Appadurai, 2004), which indicates the ability to represent the future, navigate the dense combination of nodes and pathways, set goals and make plans to reach them. This concept has been widely applied in educational research to understand students' aspirations, especially those from disadvantaged backgrounds. Indeed, social and material resources significantly influence the development of career aspirations (Mazenod et al., 2019), underscoring the need to understand how poverty traps can manifest as constraints internal to individuals, such as impaired agency or lack of hope, leading to a compromised capacity to aspire.

Although the role of the capacity to aspire has been recognized as pivotal in enabling aspirations (Bok, 2010), a clear and comprehensive operational definition of the concept, as well as its measurement, remains elusive in the existing literature. The present work aims to address this gap by proposing an operationalization of students' capacity to aspire and identifying the proper dimensions and indicators to obtain an appropriate assessment. To ensure the rigor of the research, the study employs both exploratory and confirmatory factor analysis to explore and validate the underlying factorial structure.

The remainder of the paper is structured as follows: Section 2 provides a theoretical definition of the capacity to aspire and delineates its hypothesized theoretical dimensions; Section 3 describes the data collection procedure and the adopted measures; Section 4 reports the main results obtained from statistical analysis; and Section 5 includes a discussion of the results and some conclusions.

2. Theoretical definition of capacity to aspire

Appadurai (2004) describes the capacity to aspire as the ability to read "a map of a journey into the future" (p. 76). According to this metaphor, aspiration development involves a navigational capacity in the possible future pathways, where factors such as socio-economic background can affect the destinations (the goals) and the details on the map (how to reach those goals).

Appadurai's concept applies to several contexts, including high school students' career aspirations (Bok, 2010; Mazenod et al., 2019). Various psychological factors contribute to aspiration development, reflecting what Appadurai has defined as the "capacity to aspire". First, Appadurai's framework underscores the importance of being future-oriented, which relates to how individuals prepare themselves for the future. This attitude aligns with the notion of *future concern* (Savickas and Porfeli, 2012), defined as the extent to which an individual is oriented to and involved in preparing for the future. In the educational context, concern about the future helps students look ahead and prepare for possible academic and working opportunities. However, the concern is not enough; active *exploration* of various career alternatives is crucial for setting goals and deeply evaluating current choices (Luyckx et al., 2005). Indeed, considering different pathways helps students make informed decisions about their future. *Planning* in detail when, where, and how to act to achieve one's goals and preparing for potential obstacles represent the next steps to turn intentions into actions (Sniehotta et al., 2005), ensuring that concrete actions and strategies support aspirations.

Moreover, the literature highlights that people's effort to navigate the dense combination of pathways and reach their goals is also related to their *locus of control* and *agency* (see, among others, Rottinghaus et al., 2012; Pham and Murray, 2019). The locus of control refers to a generalized attitude, belief, or expectancy regarding the nature of the causal relationship between one's behavior and its consequences (Rotter, 1966). The agency consists of the human capability to influence one's functioning and the course of events through intentional actions (Bandura, 2006), involving intentionality, forethought, self-reactiveness, and self-reflectiveness. Students with an internal locus of control believe that their efforts can lead to desired results and thus are more likely to take the initiative and persist in facing challenges. A high level of agency leads to proactive steps toward one's goals, maintaining motivation, and overcoming barriers.

Finally, as strongly underlined by Appadurai, aspirations are "always formed in the thick of social life" (Appadurai, 2004, p.67). Thus, the capacity to aspire also reflects the ability to collectively voice their needs and visions for a future with more favorable social, cultural, and economic circumstances. This facet ties into Bandura's (2000) concept of *collective efficacy*, defined as the individual's perception of the ability of their collective to be effective in achieving results as an outcome of group action. In educational settings, students who perceive strong collective efficacy among peers and teachers are more likely to feel supported and confident in their pursuits, significantly improving their capacity to aspire (Kellett et al., 2009; Cheung et al., 2013).

In sum, high levels of capacity to aspire should be reflected in individuals' (i) concern about their future (*Future concern*), (ii) belief in having control over their future (*Internal locus of control*), (iii) active exploration in breadth and depth of different educational and/or working alternatives (*Exploration*), (iv) belief in their ability to achieve the desired goals (*Agency*), (v) action planning to reach one's goals and defining coping strategies for potential obstacles (*Planning*), and (vi) confidence in their generation's capacity to change living conditions and achieve educational and/or working goals (*Collective efficacy*). These elements represent the six theorized dimensions of the "capacity to aspire" construct.

It is worth noting that the proposed dimensions of the capacity to aspire construct appear to share similarities with some aspects of the psychological capital (PsyCap) framework. PsyCap refers to a positive psychological state characterized by self-efficacy, optimism, hope, and resilience; this concept was first proposed by Luthans, Youssef, and Avolio (2007) and then adapted in the academic context by Robusto et al. (2019). In a similar way, a positive future orientation has also proved to be relevant for educational commitment and career planning (Andre, 2018). However, despite PsyCap being a prominent reference in the literature on aspiration development, it can be clearly distinguished from capacity to aspire that mainly consists of self-regulatory competencies that individuals use to guide their behavior in response to contextual opportunities. Consequently, the theoretical dimensions of the capacity to aspire construct align more closely with the career adaptability model proposed by Savickas and Porfeli (2012). Furthermore, the proposed theoretical definition highlights the importance of a social dimension in aspiration development, described in terms of collective efficacy. Indeed, the perception of the ability of their collective (herein, the one's generation) to be effective in achieving a desired outcome (herein, educational and/or working goals) influences individual orientation and effort in pursuing those goals.

3. Data collection and measures

The study involves a convenience sample of students aged 16 to 19, attending the third, fourth, and fifth years of upper secondary school in Naples, Italy. After a data cleaning process, n=191 students were considered for the analysis. Most participants are female (58%) and came from technical-professional schools (49%) followed by high schools (36%).

The six dimensions of the capacity to aspire were measured using a set of indicators with a 5-point Likert response scale. Specifically, for the dimensions of future concern, locus of control, exploration, agency, and planning, the scale ranges from 1 = Strongly disagree to 5 = Strongly agree. Conversely, the collective efficacy item response scale ranges from 1 = Not at all able to 5 =

Completely able. The initial pool of developed indicators[1] included three items for future concern (adapted from Savickas et al., 2012; item example: "I am aware that today's choices shape my future"), two items for internal locus of control (adapted from Farma et al., 2000; item example: "I believe a person can truly be the author of his own destiny"), six items for exploration (adapted from Porfeli et al., 2011; item example: "I think about how I could fit into many different educational and/or working paths"), four items for agency (adapted from Rottinghaus et al., 2012; item example: "I feel able to make a correct educational and/or job choice for the future"), four items for action and coping planning (adapted from Lo Presti et al., 2012; item example: "I know exactly what I have to do to achieve my educational and/or working goals"), and four completely new items for individual perception of collective efficacy among one's generation (item example: "People belonging to my generation can act to change their living conditions and have a better future"). Hence, a total of 23 items were considered to obtain a measure of the "capacity to aspire" construct.

4. Statistical analysis

The capacity to aspire has been conceived as a multidimensional latent variable measured by Likert-type items. Accordingly, exploratory and confirmatory factor analyses (Thompson, 2004) were employed to explore and validate the underlying factorial structure. The original sample was split into two nearly equal groups of n_1=100 and n_2=91 observations. Exploratory factor analysis (EFA) was carried out on the first sub-sample to identify the underlying dimensions within the considered items; then, the emerged factorial structure was validated on the second sub-sample by exploiting the confirmatory factor analysis (CFA). To ascertain the sampling adequacy for EFA, the *Kaiser-Meyer-Olkin* (KMO) test and *Bartlett's test of Sphericity* were conducted, reporting a satisfactory correlation between variables: the KMO value was equal to 0.76, and *Bartlett's test of Sphericity* was significant at α=0.001. At the first round of EFA, the number of factors was selected according to the *parallel analysis* suggesting four latent dimensions. Items with very low communality, low factor loadings, or high cross-loadings were removed, one at a time, within an iterative process. The final solution, depicted in Figure 2, includes 13 of the original 23 items, measuring four latent dimensions and explaining 55% of the variability. The *maximum likelihood* method was used to estimate model parameters, and an *oblimin* rotation was applied to improve factor interpretation. The list of the selected items is in the Appendix.

Results from the CFA based on the emerged factorial structure with the selected 13 items confirmed the adequacy of the measurement model. Most items had factor loading values greater than 0.7, and fit measures support the overall good fit of the model (CFI = 0.977; TLI = 0.970; RMSEA = 0.042; SRMR = 0.066). In addition, a comparison between the four first-order correlated factors model and a simpler overall second-order factor model showed no significant difference ($\Delta\chi^2 = 0.10$; *p*-value$= 0.95$), suggesting that the four first-order latent variables represent facets of a single general factor herein named "capacity to aspire".

[1] The set of considered indicators is available upon request to the first author.

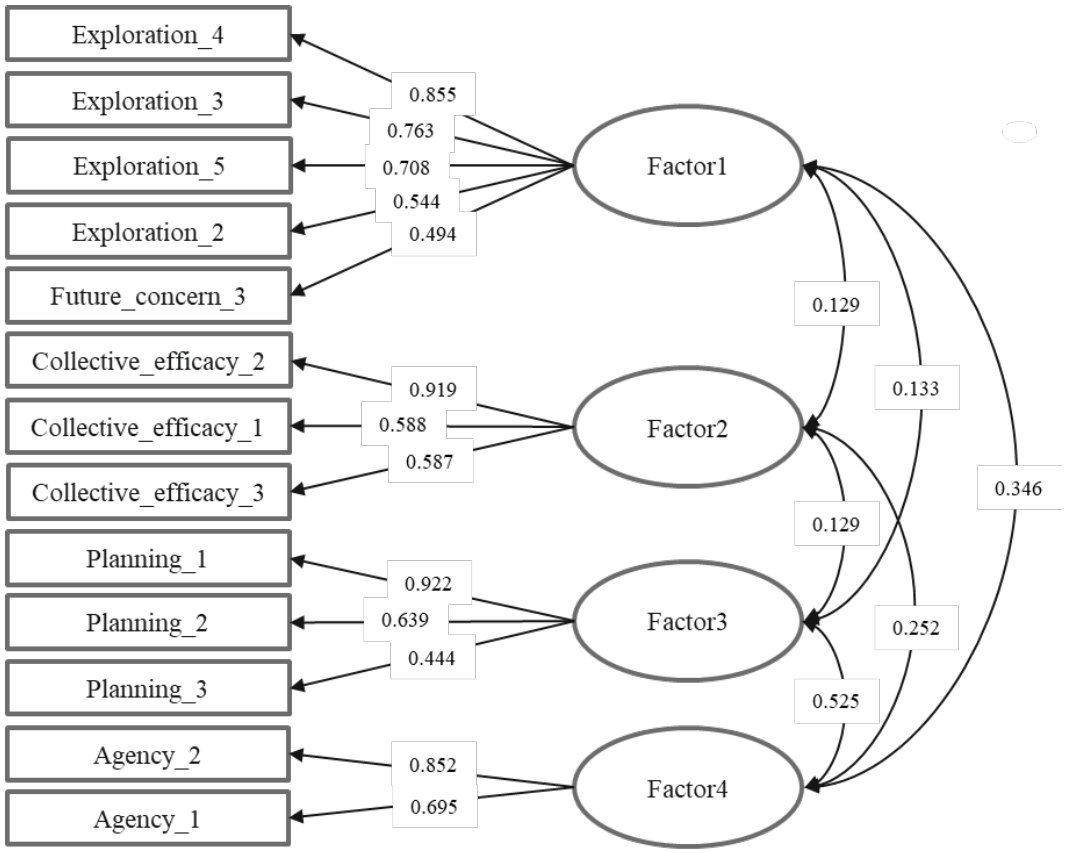


Figure 2 – Diagram of the final EFA model

5. Discussion and conclusion

The present work focused on developing and validating a comprehensive measure to assess high school students' capacity to aspire. Based on the EFA and according to the results obtained from CFA, a set of indicators was selected and tested for their adequacy. The findings revealed a factorial structure with four main dimensions of the capacity to aspire and a second-order factor reflecting the general concept. The four specific dimensions are: Future concern and exploration (Factor1), Collective efficacy (Factor2), Planning (Factor3), and Agency (Factor4). The internal locus of control dimension was completely discarded during the item selection, highlighting its non-significant contribution to the capacity to aspire measurement.

A significant limitation of this study is the small and non-probabilistic nature of the sample, which prevents the generalization of results to the entire population. Nevertheless, the present research offers valuable insights, as it represents an initial attempt to address a gap in the literature by providing an operationalization of high school students' capacity to aspire and developing a set of indicators for its measurement.

Future research could test the emerged factorial structure on a larger sample and explore the measurement invariance of the proposed scale according to some socio-demographic variables (e.g., gender). Another future study route could investigate the relationship between students' socio-economic backgrounds and career aspirations, mediated by the capacity to aspire. Results from such studies could inform policymakers about possible interventions to foster material resources and cultural ability, thereby expanding aspirational capacity among disadvantaged groups.

Acknowledgment

Rosa Fabbricatore acknowledges the financial support provided by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and

Sustainable project (GRINS PE00000018 - CUP E63C22002140007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

This work was also supported by the University Research Funding Program (FRA) 2022 of the University of Naples Federico II, with the contribution of the Compagnia San Paolo (Title of the project: "Measuring and Mapping Educational Poverty"; principal investigator: Cristina Davino).

References

Andre, L., van Vianen, A.E.M., Peetsma, T.T.D., Oort, F.J. (2018). Motivational power of future time perspective: Meta-analyses in education, work, and health. *PLoS One*, **13**(1), e0190492.

Appadurai, A. (2004). The capacity to aspire: culture and terms of recognition, in *Culture and Public Action*, eds. R. Vijayendra and M. Walton, Stanford University Press, Stanford, (CA), pp. 59-84.

Bandura, A. (2000). Exercise of human agency through collective efficacy. *Current Directions in Psychological Science*, **9**(3), pp. 75-78.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, **1**(2), pp. 164-180.

Bok, J. (2010). The capacity to aspire to higher education: 'It's like making them do a play without a script'. *Critical Studies in Education*, **51**(2), pp. 163-178.

Cheung, F.M., Wan, S.L., Fan, W., Leong, F., Mok, P.C. (2013). Collective contributions to career efficacy in adolescents: A cross-cultural study. *Journal of Vocational Behavior*, **83**(3), pp. 237-244.

Farma, T., Cortivonis, I. (2000). Un questionario sul "locus of control": Suo utilizzo nel contesto italiano (A questionnaire on the locus of control: Its use in the Italian context). *Ricerca in Psicoterapia*, **3**(2-3), pp. 147-155.

Kellett, J.B., Humphrey, R.H., Sleeth, R.G. (2009). Career development, collective efficacy, and individual task performance. *Career Development International*, **14**(6), pp. 534-546.

Lo Presti, A., Pace, F., Mondo, M., Nota, L., Casarubia, P., Ferrari, L., Betz, N.E. (2013). An examination of the structure of the career decision self-efficacy scale (short form) among Italian high school students. *Journal of Career Assessment*, **21**(2), pp. 337-347.

Luyckx, K., Goossens, L., Soenens, B., Beyers, W., Vansteenkiste, M. (2005). Identity statuses based on 4 rather than 2 identity dimensions: Extending and refining Marcia's paradigm. *Journal of Youth and Adolescence*, **34**(6), pp. 605-618.

Luthans, F., Youssef, C.M., Avolio, B.J. (2007). *Psychological Capital: Developing the Human Competitive Edge*. Oxford University Press, Oxford, (UK).

Mazenod, A., Hodgen, J., Francis, B., Taylor, B., Tereshchenko, A. (2019). Students' university aspirations and attainment grouping in secondary schools. *Higher Education*, **78**, pp. 511-527.

Pham, Y.K., Murray, C. (2019). Career locus of control and school-and career-related adjustment among high-need youth with and without disabilities. *Journal of Career Development*, **46**(5), pp. 502-515.

Porfeli, E.J., Lee, B., Vondracek, F.W., Weigold, I.K. (2011). A multi-dimensional measure of vocational identity status. *Journal of Adolescence*, **34**(5), pp. 853-871.

Robusto, E., Maeran, R., Colledani, D., Anselmi, P., Scioni, M. (2019). Development of two scales for measuring academic psychological capital and locus of control in fresh graduates. *Statistica Applicata – Italian Journal of Applied Statistics,* **31**(1), pp. 13-28.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, **80**(1), pp. 1-28.

Rottinghaus, P.J., Buelow, K.L., Matyja, A., Schneider, M.R. (2012). The career futures inventory–revised: Measuring dimensions of career adaptability. *Journal of Career Assessment*, **20**(2), pp. 123-139.

Savickas, M.L., Porfeli, E.J. (2012). Career adapt-abilities scale: Construction, reliability, and measurement equivalence across 13 countries. *Journal of Vocational Behavior*, **80**(3), pp. 661-673.

Sniehotta, F.F., Schwarzer, R., Scholz, U., Schüz, B. (2005). Action planning and coping planning for long-term lifestyle change: theory and assessment. *European Journal of Social Psychology*, **35**(4), pp. 565-576.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. American Psychological Association, Washington, (DC).

Appendix

Selected items to measure the capacity to aspire:

Exploration_2:	I think about how I could fit into many different educational and/or working paths
Exploration_3:	I gather information about different educational and/or working paths that might interest me
Exploration_4:	I carefully look into the requirements for the educational and/or working path I am most interested in
Exploration_5:	I try to understand how to improve my chances of pursuing the educational and/or working path I have chosen
Future_concern_3:	I am aware that today's choices shape my future
Collective_efficacy_1:	People belonging to my generation can imagine different educational and/or working options to choose from
Collective_efficacy_2:	People belonging to my generation can achieve their educational and/or working goals
Collective_efficacy_3:	People belonging to my generation can act to change their living conditions and have a better future
Planning_1:	I know exactly what I have to do to achieve my educational and/or working goals
Planning_2:	I have already planned a series of actions to achieve my educational and/or working goals (e.g., attending training courses, saving money, contacting organizations or qualified individuals)
Planning_3:	I know how to deal with any potential challenges I may encounter during my future educational and/or working path
Agency_1:	I feel able to make a correct educational and/or job choice for the future
Agency_2:	I believe I am able to successfully handle the educational and/or working transitions that lie ahead

doi.org/10.26398/asaproc.0011

Improving online misogyny detection: The role of human annotation and explanation-guided learning

Lara Fontanella[a], Alex Cucco[a,d], Emiliano del Gobbo[b], Elisa Ignazzi[c], Sara Fontanella[d]

[a] Department of Legal and Social Sciences, G. d'Annunzio University, Chieti-Pescara, Italy
[b] Department of Economics, Management and Territory, University of Foggia, Italy
[c] Department of Neuroscience, Imaging and Clinical Sciences, G. d'Annunzio University, Chieti-Pescara, Italy
[d] National Heart and Lung Institute, Imperial College London, UK

1. Introduction

In today's world, despite decades of advocating for equal rights, women continue to face disproportionate hostility compared to men in many situations. Misogyny is the term used to describe the widespread antagonism and aversion towards women and girls. It originates from the Ancient Greek word "μισογυνία", which translates to hatred of women and can be broadly defined as the cultural mindset of directing hatred towards females solely because of their gender (Johnson, 2000).

The advent of social media has been widely acknowledged as a catalyst for the proliferation of misogynistic content. Indeed, although digital platforms have increased the visibility of women's voices, they have also notoriously enabled the dissemination of misogynistic content (Amnesty International Italia, 2020; European Union Agency for Fundamental Rights, 2023). However, due to the vast quantities of data generated by social media users, human monitoring of their content is unfeasible. Therefore, in the wider domain of hate speech investigation, there has been a significant focus on developing techniques to automatically detect offensive material on the internet (see Tontodimamma et al., 2021, for a review). Hate speech detection can be regarded as a challenging classification task commonly tackled using supervised models. In this context, some studies primarily concentrate on automatic misogyny identification (AMI) and employ a range of methodologies (Fontanella et al., 2024, for a review). Most early AMI research predominantly relied on machine learning (ML) classifiers, with SVM being the most popular (Canós, 2018), followed by logistic regression (Shushkevich and Cardiff, 2018), naïve Bayes (Nwesri et al., 2021), and random forest (Liu et al., 2018).

Conventional ML models need meticulous engineering and extensive domain knowledge to create a feature extractor that converts the raw data into a suitable internal representation or feature vector that the learning algorithm can use to categorise patterns in the textual input. On the other hand, deep learning (DL) models allow the system to bypass the need for manual feature engineering by learning relevant representations directly from the data (LeCun et al., 2015). Deep learning methods have shown significant success in AMI, particularly for their strong performance. Notably, transformer-based natural language process (NLP) models have become the standard for misogyny detection (see Rodríguez et al., 2023, Tri Wibowo et al., 2023, among others).

More recently, large language models such as GPT (generative pre-trained transformer) have further demonstrated their efficacy for AMI purposes (Pérez- Almendros and Camacho-Collados, 2024, Sultana and Kali, 2024). Appropriate models for identifying misogynistic discourse offer scalability and reduce the mental burden on content moderators. However, these models often rely on complex techniques that are difficult for humans to understand, posing a challenge in providing clear and ethical justifications for their decisions. Black-box models, which lack transparency, are unsuitable for moderating online comments, as both users and moderators are sceptical of decisions

they cannot comprehend. Offering explanations helps build trust and increases the acceptability of machine-learning classifiers. This need for explainability has given rise to explainable AI (XAI) (Adadi and Berrada, 2018) which has become essential for understanding deep neural networks. In AMI tasks, it has been shown how post-hoc explanation methods (Arnob et al., 2023, Attanasio et al., 2022), such as Shapley additive explanations (Lundberg and Lee, 2017), sampling-and-occlusion (Jin et al., 2020) and local interpretable model-agnostic explanations (Ribeiro et al., 2016), provide plausible and faithful explanations. Post-hoc explanation methods are used to interpret and understand the decisions made by DL models after they have been trained and have made predictions, and human involvement is typically limited to interpreting the results of the explanation methods.

However, such approaches often lack certainty due to the absence of ground truth for machine decision-making. On the other hand, human involvement is central to explanation guided learning (EGL) (see Gao et al., 2024, for a review). During training, EGL models are guided by human-provided explanations or criteria to focus on certain aspects of the input that are deemed important for decision-making. In NLP tasks, the explanation of natural language data is typically represented by *rationales*, usually obtained by human annotation, that highlight the most significant parts of the textual data for making specific task predictions. During model training, it is common practice to optimise both the explanation loss and the conventional prediction loss simultaneously (Gao et al., 2024). The EGL approach, incorporating the generation of explanations into the model training process, ensures that the model's decision-making process is transparent and aligned with human expectations from the beginning. Despite the challenges in aligning model-generated explanations with human annotations and evaluating their impact on prediction accuracy, explanation quality, and model generalisability, EGL techniques show promise for AMI tasks.

In this work, we aim to explore how EGL can be effectively applied to identify the rationales supporting the classification of short texts, such as online comments. Understanding the justification behind classifying a text is crucial for developing accurate and interpretable models. By leveraging human-annotated data, we seek to enhance both the explainability and fairness of our classification model. The specific model utilised in this study is detailed in Section 2, while Section 3 presents the results of applying EGL for automatic misogyny identification in Italian online comments.

2.The ExPred model

In this section, we briefly introduce ExPred (Zhang et al., 2021), a model that processes textual input for tasks like rationale-based classification, where a rationale is a concise and crucial portion of the text that justifies the prediction. ExPred is an EGL model that falls into the framework of so-called explain-then-predict models. Such models use a two-stage process: explanation and prediction. In the explanation phase, a model learns to extract a rationale from input text. Then, in the prediction phase, another independent model predicts the task output solely based on the extractive explanation. Specifically, the ExPred algorithm (Figure 1) learns to produce explanations under the supervision of both task-specific and rationale-based criteria during the explanation phase, adopting a multi-task machine learning approach. The ExPred input is a sequence of words represented as vectors in an embedding space, with each word labelled to indicate whether it is part of a rationale. In the two-part architecture, the explanation generator works alongside an auxiliary task predictor, both sharing a BERT-based encoder. In this scenario, the auxiliary task involves classifying the entire text. For the auxiliary task, a multi-layer perceptron (MLP) decoder is used to map the BERT-encoded input to the task prediction. The decoder for the explanation network uses a binary classification setup over the input sequence of words to identify rationale tokens.

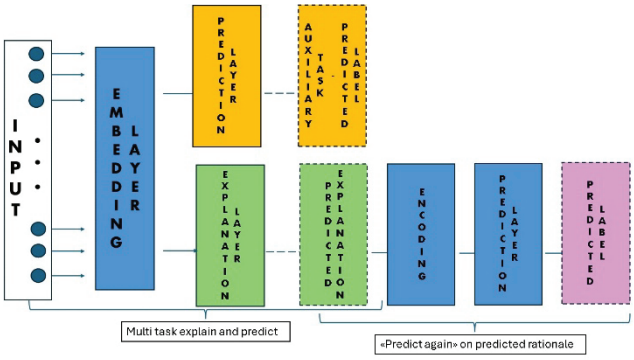


Figure 1: Summary of ExPred. The method generates explanations guided by task objectives and the explanation generation process itself. The auxiliary task is identical to the final task.

A gated recurrent unit is used to capture the sequential context of the words, and pooling techniques are employed to create comprehensive word representations. Finally, a word-wise MLP makes the binary decision for each word, determining the set of words that constitute the explanation. The model is trained optimising both the task prediction and explanation generation. The loss function $L_{loss} = L_{task} + \lambda L_{exp}$ balances the accuracy of the task prediction L_{task} and the quality of the explanation L_{exp} according to a parameter λ that regulates the importance of loss between task and explanation. After training the explanation model, in the prediction phase, a separate model is trained specifically for making predictions based only on the rationale extracted in the previous phase to predict the final outcome. BERT is employed as the model for making the final task prediction based on the masked input. The training of the prediction network considers only the instances where the auxiliary task prediction (from the first phase) matches the actual task label. The model is also validated using masked inputs but unlike during training, no input instances are excluded during validation based on the auxiliary model's predictions.

3. Automatic misogyny identification in Italian online comments

3.1 Corpus collection

The analysed corpus is a subset of a broader collection of Twitter, Facebook, and Instagram comments that were chosen and collected via various techniques. The Twitter API was used to retrieve real-time tweets from 40 Italian political activists, feminists, journalists, influencers, and female TV celebrities, indexed by their first name and last name. In addition, we selected 223 Facebook posts related to Italian women and 63 Instagram posts from influencers. The relevant comments were obtained by downloading them through exportcomments.com. After analysing the entire corpus, we selected a sample of 13,500 comments for manual annotation. These comments varied among social media platforms and specifically mentioned women.

3.2 Corpus annotation

Ten annotators, who were Master's students specialising in social research, were trained and divided into five groups, each consisting of a woman and a man. Each group was assigned a portion of the dataset for annotation. Within each group, the two annotators conducted the annotation independently using a binary scheme for misogyny detection. Seminal works (Zaidan et al., 2008) have collected annotator rationales by asking human annotators to highlight parts of a text to justify "why" that text should receive a certain label. Following this approach, if a comment was deemed misogynistic, the annotators identified and selected the specific part of the text that demonstrated misogynistic content, thus obtaining snippet-level rationales consisting of multiple consecutive words. It is important to note that an annotated rationale can consist of multiple text spans that together form one comprehensive rationale explaining the decision.

3.3 Training, validation and test sets

Due to potential inconsistencies in the human annotation process, particularly in the classification and identification of rationales, the division between training, validation, and test sets was carried out meticulously. For training and validating the model, we relied solely on posts that were consistently classified by both annotators as either misogynistic or non-misogynistic. The training set was formed of 5996 comments, of which 2999 were misogynistic and 2997 were nonmisogynistic. The same approach was applied to construct the validation set, which included 500 misogynistic comments and 498 non-misogynistic comments. The test set, however, consisted of 1500 comments that both annotators recognised as non-misogynistic and 1498 comments that only one annotator identified as misogynistic. The rationale behind this division is to train the model on data where annotators agree and to evaluate the model's performance in the context of human uncertainty. Consequently, when assessing the model's performance on the training set, qualitative evaluation of the model-predicted rationales and assigned labels should be prioritised. A general understanding of the model's overall performance, in cases of annotation agreement, can always be inferred from the validation set.

4.ExPred preliminary results

The analysis was performed using the Python code[1] provided by Zhang et al. (2021). We used an Italian version of BERT[2] as shared encoder for the explanation phase and for the prediction phase. While we maintained the network structure and parameters used in Zhang et al. (2021), the model was trained and validated using four different values for the λ parameter. The performance results, presented in Table 1, demonstrate the robustness of the model across various choices of the parameter.

	NON-MYSOGINISTIC			MYSOGINISTIC			WEIGHTED AVERAGE		
	F1-score	precision	recall	F1-score	precision	recall	F1-score	precision	recall
λ=5	0.78	0.85	0.72	0.81	0.76	0.87	0.79	0.80	0.79
λ=20	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
λ=50	0.78	0.75	0.81	0.76	0.80	0.73	0.77	0.77	0.77
λ=75	0.77	0.72	0.82	0.74	0.80	0.69	0.75	0.76	0.76

Table 1: Performance results on the validation set for different λ values.

For λ=5, Figure 2 presents the confusion matrix for the prediction task on the test set, along with the key terms from the rationales corresponding to both correctly classified and misclassified textual documents. From the top-right of the confusion matrix (false-negative), mislabelled misogynistic text can be identified, while the bottom-left (false-positive) highlights a well-known issue: the algorithmic bias in classify as misogynistic text including nouns referring to woman disregarding the context (Nozza et al., 2019).

[1] https://github.com/JoshuaGhost/expred

[2] bert-base-italian-xxl-uncased, https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

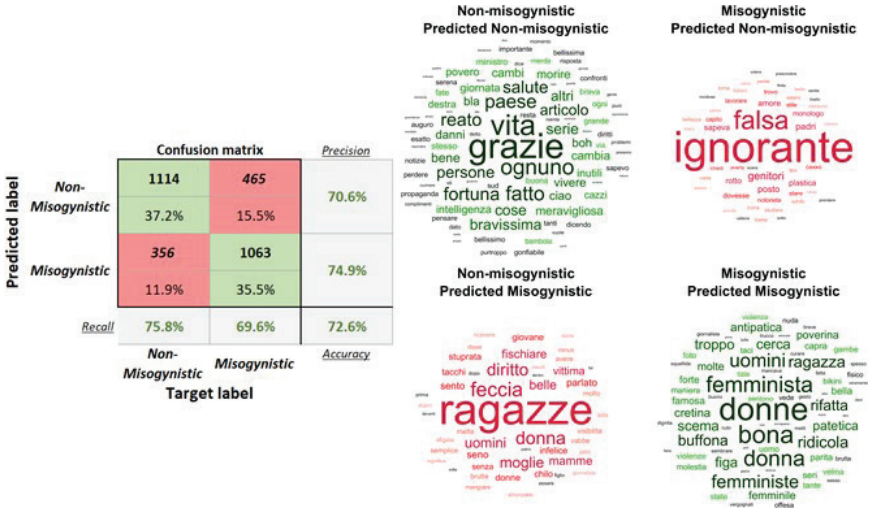


Figure 2: Confusion matrix for the test set and keywords' wordclouds - *the size of each word reflects its chi-squared keyness measure*.

5. Concluding remarks

Despite progress in misogyny detection using deep learning models, providing clear and ethical justifications for their decisions remains a challenge. This study investigates the use of EGL on a corpus of Italian social media comments to evaluate the explainability and fairness of this type of misogyny detection model, ensuring their decision-making aligns with human expectations. While EGL represents a tailored framework to incorporate human-annotated rationales, there are limitations to consider. The model utilised may be prone to error propagation. Additionally, in the real case presented, as we combined two sets of annotations, the union of the identified rationales was used, and for non-misogynistic instances, the rationale was provided as the entire text. Future studies should address the one-class intrinsic nature of the problem to further refine the model's performance.

Acknowledgement: *This research was undertaken as part of the ICOMIC (Identifying and Counteracting Online Misogyny in Cyberspace) Project funded by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021*

References

Adadi, A., Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, pp. 52138–52160.

Amnesty International Italia. (2020). *Barometro dell'odio: Sessismo da tastiera*. Retrieved from https://www.amnesty.it/barometro-dellodio-sessismo-da-tastiera/.

Arnob, S.S., Shikder, M.A., Ovey, T.A., Rhythm, E.R., Rasel, A.A. (2023). Bengali misogyny identification with deep learning and LIME. In *Proceedings of the 2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pp. 285–292.

Attanasio, G., Nozza, D., Pastor, E., Hovy, D. (2022). Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pp. 100–112

Canós, J. (2018). Misogyny identification through SVM at IberEval 2018. In *CEUR Workshop Proceedings*, 2150, pp. 229–233.

European Union Agency for Fundamental Rights. (2023). *Online Content Moderation - Current Challenges in Detecting Hate Speech*. Retrieved from https://fra.europa.eu/en/publication/2023/online-content-moderation

Fontanella, L., Chulvi, B., Ignazzi, E., Sarra, A., Tontodimamma, A. (2024). How do we study misogyny in the digital age? A systematic literature review using a computational linguistic

approach. *Humanities and Social Sciences Communications*, 11(478), pp. 1–15.
Gao, Y., Gu, S., Jiang, J., Hong, S., Yu, D., Zhao, L. (2024). Going beyond XAI: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7), pp. 1–39.
Jin, X., Wei, Z., Du, J., Xue, X., Ren, X. (2020). Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.
Johnson, A.G. (2000). *The Blackwell Dictionary of Sociology: A User's Guide to Sociological Language*. John Wiley & Sons Incorporated.
LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), pp. 436–444.
Liu, H., Chiroma, F., Cocea, M. (2018). Identification and classification of misogynous tweets using multi-classifier fusion. In *CEUR Workshop Proceedings*, 2150, pp. 268–273.
Lundberg, S.M., Lee, S. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777.
Nozza, D., Volpetti, C., Fersini, E. (2019). Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 149–155.
Nwesri, A., Wu, S., Harmain, H. (2021). Detecting misogyny in Arabic tweets. In *CEUR Workshop Proceedings*, 3159, pp. 861–866.
Pérez-Almendros, C., Camacho-Collados, J. (2024). Do large language models understand mansplaining? Well, actually... In *Proceedings of the LREC-COLING 2024 – Main Conference*, pp. 5235–5246.
Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144
Rodríguez, D.A., Diaz-Escobar, J., Díaz-Ramírez, A., Trujillo, L. (2023). Domain-adaptive pre-training on a BERT model for the automatic detection of misogynistic tweets in Spanish. *Social Network Analysis and Mining*, 13(126).
Shushkevich, E., Cardiff, J. (2018). Misogyny detection and classification in English tweets: The experience of the ITT team. In *CEUR Workshop Proceedings*, 2263.
Sultana, S., Kali, M. B. (2024). Exploring ChatGPT for identifying sexism in the communication of software developers. In *Proceedings of the 17th International Conference on Pervasive Technologies Related to Assistive Environments*, pp. 400–403.
Tontodimamma, A., Nissi, E., Sarra, A., Fontanella, L. (2021). Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126, pp. 69–81.
Tri Wibowo, B., Nurjanah, D., Nurrahmi, H. (2023). Identification of misogyny on social media in Indonesian using bidirectional encoder representations from transformers (BERT). In *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 401–406.
Zaidan, O.F., Eisner, J., Piatko, C. (2008). Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NeurIPS2008 Workshop on Cost Sensitive Learning*.
Zhang, Z., Rudra, K., Anand, A. (2021). Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 418–426.

doi.org/10.26398/asaproc.00112

The evolution of Explainable Artificial Intelligence (XAI): A preliminary systematic literature review

Agostino Gnasso, Massimo Aria
Department of Economics and Statistics, University of Naples Federico II.

1. Introduction

The global increase in research on explainable artificial intelligence (XAI) over the past decade has led to the development of numerous domain-specific methods for interpreting machine learning (ML) models and generating human-understandable explanations (Abusitta et al., 2024; Aria et al., 2021). This growth is largely driven by the widespread adoption of ML, particularly deep learning, across various sectors such as e-commerce, healthcare, and finance, where transparency in AI systems is essential (Adadi and Berrada, 2018). XAI is particularly crucial in critical domains like healthcare, where it aids in interpreting ML models for disease prediction and diagnosis, and in finance, where it helps explain credit scoring models and fraud detection systems, ensuring regulatory compliance and customer trust (Whig et al., 2023; Bussmann et al., 2021). Previous studies highlight the need to organize the literature on AI explainability across various subdomains (Burkart and Huber, 2021; Molnar et al., 2020; Guidotti et al., 2018). Most XAI methods focus on interpreting the entire ML process, with explanations taking forms such as rules, numerical data, or visuals, evaluated using metrics often derived from human-computer interaction (HCI) (Aria et al., 2024). Building on prior efforts, this article offers a systematic, transparent, and reproducible bibliometric review (Donthu et al., 2021) of global XAI research. Through the systematic literature review (SLR) of 6,148 documents from the Web of Science database, this study provides an in-depth analysis of the current state, historical evolution, and future directions of XAI research, including trends in scientific activity, key publications, author contributions, and international collaboration, using the Bibliometrix R package (Aria and Cuccurullo, 2017).

2. Methods and results

The selection process was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). Additionally, a bibliometric analysis was conducted for the period from January 2010 to December 2023, utilizing the open-source R package Bibliometrix (Aria and Cuccurullo, 2017) to analyze documents retrieved from the Web of Science (WoS) database, managed by Clarivate Analytics. The query was designed to identify a comprehensive range of terminologies and acronyms that have emerged in both theoretical and practical contexts. This includes terms such as "explainable AI," "interpretable machine learning," and expressions related to "black-box" models. The selection query was formulated as follows: “Database: Web of Science Core Collection. TOPIC: (“explainable machine learning” OR “explainable artificial intelligence” OR “explainable ai” OR “interpret* machine learning” OR “interpret* artificial intelligence” OR “interpretable ai” OR (“explainab*” AND “XAI”) OR (“explainab*” AND “XML”) OR (“explainab*” AND “black box*”) OR (“interpret*” AND “black box*”) OR (“XAI” AND “black box*”) OR (“XML” AND “black box*”)) Refined by: DOCUMENT TYPES: (ARTICLE OR PROCEED- ING PAPER). Languages = English. Indexes = SCI-EXPANDED, SSCI. Time = 2010–2023.” The year 2010 was chosen as the starting point for the analysis because it marks the first year in which consistent metadata for documents within the WoS database was available. Selecting 2010 allows for the capture of a complete decade-plus span of research, which provides sufficient data to identify trends, significant developments, and the evolution of the field over time. **Table 1** provides a summary of bibliometric data on research

outputs related to explainable artificial intelligence (XAI) from 2010 to 2023. The collection includes a total of 6,148 documents sourced from 1,731 different journals, books, and other sources, with an average of 16.86 citations per document. These documents collectively reference over 239,000 other works. Moreover, the collection includes contributions from 22,846 authors. Only 228 documents are single-authored, showing a strong trend toward collaborative research. On average, each document has about 5 co-authors. Additionally, 33.23% of the documents involve international co-authorship, reflecting a significant level of global collaboration in XAI research. The documents include a substantial number of keywords, with 7,977 entries under Keywords Plus and 14,189 entries under Author's Keywords, indicating a rich variety of topics and terms used in this field. Keywords Plus in the WoS database are automatically generated terms derived from the titles of cited references in an article, enhancing discoverability by providing additional relevant indexing beyond the author's chosen keywords.

MAIN INFORMATION ABOUT DATA	
Timespan	2010:2023
Sources (Journals, Books, etc)	1731
Documents	6148
Average citations per documents	16,86
References	239420
DOCUMENT TYPES	
Article	5917
Early Access	189
Proceedings Paper	42
DOCUMENT CONTENTS	
Keywords Plus (ID)	7977
Author's Keywords (DE)	14189
AUTHORS	
Authors	22846
Authors of single-authored documents	228
AUTHORS COLLABORATION	
Single-authored documents	248
Co-Authors per Documents	5,07
International co-authorships %	33,23

Table 1 *Main Information About Data*

Figure 1 represents the annual scientific production, specifically the number of articles published each year from 2010 to 2023. The graph shows a dramatic increase in the number of articles over time, particularly from around 2018 onwards. The relatively flat trend from 2010 to 2017 suggests a slow initial growth in scientific publications related to the topic of interest, followed by an exponential surge starting in 2018. This sharp increase in recent years indicates a rapidly growing interest and research activity in the field, likely driven by advancements in technology, the rise in practical applications, and the increasing recognition of the importance of the subject matter. The annual growth rate of 47.03% further emphasizes the accelerating expansion of research in this area, reflecting the field's burgeoning development and its expanding influence across various sectors.

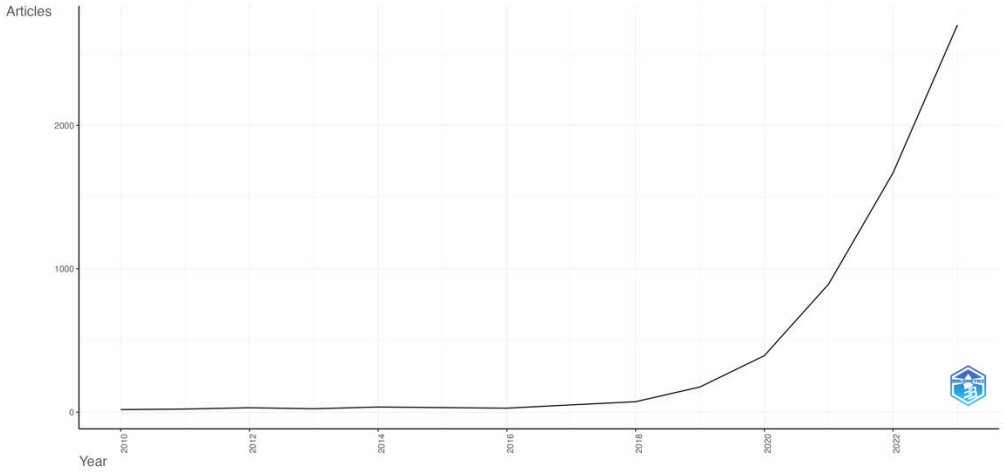


Figure 1 *Annual Scientific Production: 47.03 %*

Figure 2 represents a co-occurrence network of keywords in the field of research related to explainable artificial intelligence (XAI). In such a network, nodes represent keywords, and the size of each node reflects the frequency of that keyword's occurrence within the dataset. The lines (edges) connecting the nodes indicate co-occurrences, showing how often these keywords appear together in the same documents. The thickness of the edges signifies the strength of the co-occurrence relationship. The largest nodes, such as "classification," "model," "prediction," and "neural networks," are central to the network, indicating they are frequently occurring and are highly connected with other keywords. This suggests that these are the core themes around which much of the research revolves. The network has distinct groups of keywords, each likely representing a subdomain or a specific aspect of the research field. The green group (centered around "black-box" and "artificial intelligence") might represent research focused on the challenges and solutions associated with explaining complex, non-transparent models (black-box models). The red group (centered around "health," "disease," "mortality," etc.) likely pertains to research in the healthcare domain, where XAI is applied to interpret models related to health outcomes. The purple group seems to focus on healthcare as well, with keywords like "cancer" and "diagnosis," possibly highlighting XAI's role in medical diagnostics.

Figure 3 and Figure 4 represent a thematic analysis of research in explainable artificial intelligence (XAI), showing both a thematic network and a thematic map. The thematic network illustrates the relationships between different research themes in the field. Central themes such as "classification," "prediction," and "model" are depicted in larger, central nodes, indicating their foundational role in the research landscape. Surrounding these are groups of related concepts, such as "neural networks," "black-box," and "artificial intelligence," which connect with other relevant themes. The network is divided into groups, with some focused on specific applications (e.g., healthcare with terms like "diagnosis" and "risk") and others on technical aspects (e.g., "algorithm," "optimization"). The connectivity of these groups indicates a high level of interdisciplinary research and the integration of various subfields within XAI. The thematic map further categorizes these themes into four quadrants based on their centrality (relevance) and density (development). Basic Themes (bottom right

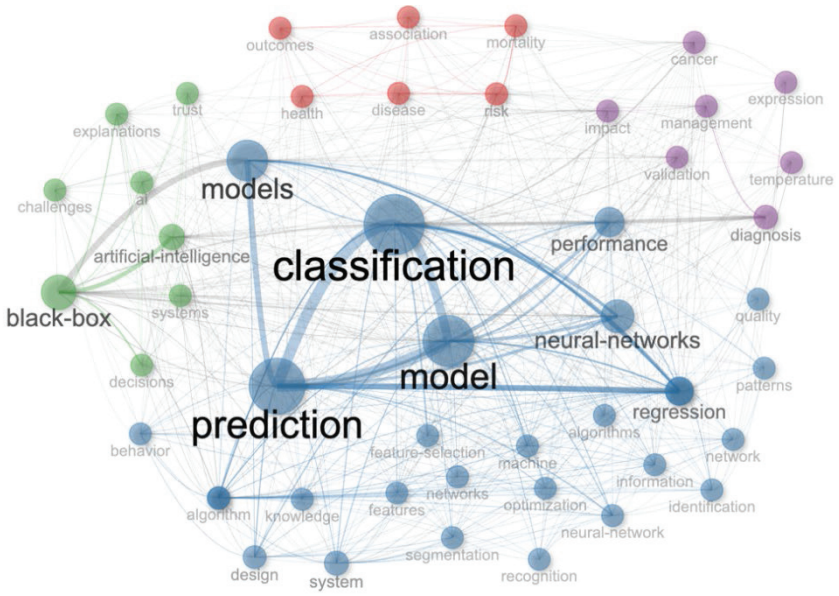


Figure 2 *Co-occurrence Network*

quadrant): The green group, which includes "classification," "prediction," and "model," represents fundamental themes that are central to the research field but less developed. These are the core topics that form the foundation of XAI research. Motor Themes (top right quadrant): The red group, featuring "models," "neural networks," and "black-box," are highly central and well-developed, indicating they are both influential and actively researched. These themes are driving the field forward, representing cutting-edge topics with significant momentum. Niche Themes (top left quadrant): The grey group, which includes terms like "mortality," "outcomes," and "association," represents specialized areas with high density but lower centrality. These themes are well-developed but may be more isolated or focused on specific, perhaps niche, applications within the broader field. Emerging or Declining Themes (bottom left quadrant): The blue group, with terms such as "risk," "impact," and "validation," represents areas that are currently less developed and may either be emerging as new areas of interest or declining in relevance.

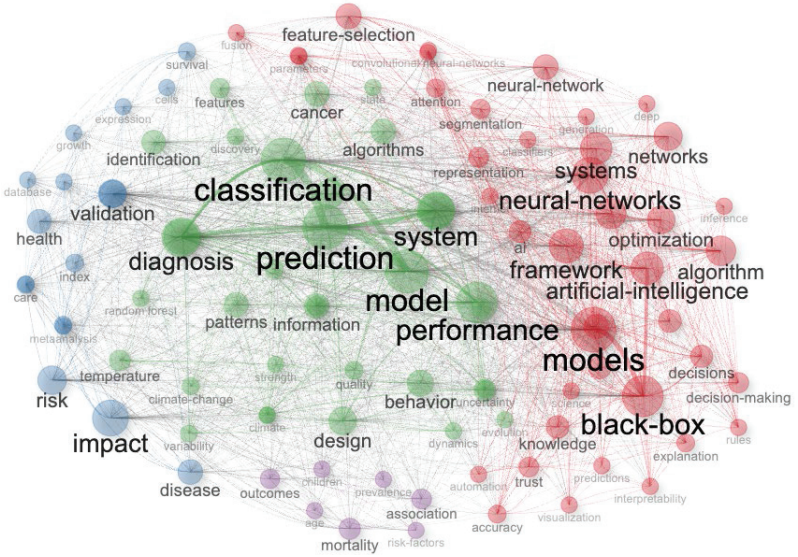


Figure 3 *Thematic Map Network*

These themes are not yet central to the field but could represent future trends or fading topics. The combination of the thematic network and thematic map provides a comprehensive view of the research landscape in XAI. Core themes like "classification" and "prediction" form the backbone of the field, while dynamic and evolving areas like "neural networks" and "black-box" models are at the forefront of ongoing research. Meanwhile, specialized and potentially emerging themes highlight the diversity and complexity of research within this domain. This analysis underscores the multidisciplinary nature of XAI and points to both established and evolving areas of interest in the field.

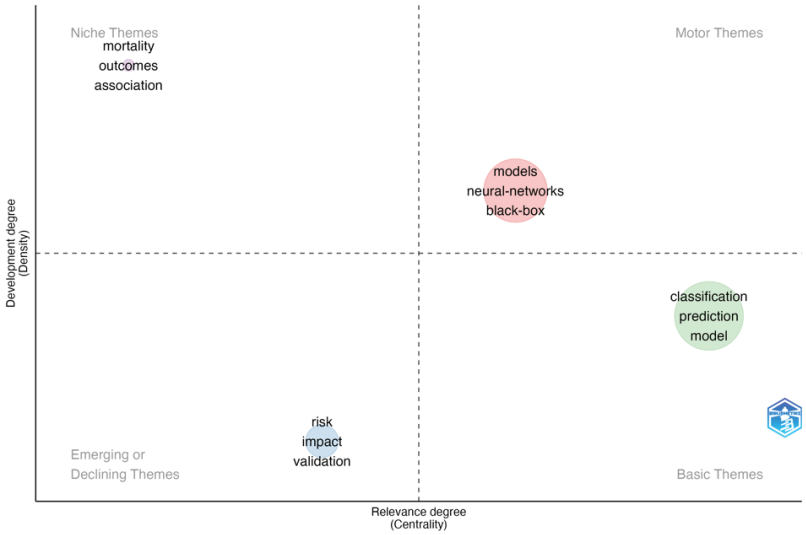


Figure 4 *Thematic Map*

3. Conclusions

This study provides a preliminary analysis of the current research landscape in Explainable Artificial Intelligence (XAI), highlighting key themes and emerging trends in the field. By examining the evolution of scientific production, we have identified a significant increase in research activity, particularly in recent years, underscoring the growing importance of XAI in both academic and practical contexts. The analysis of keyword co-occurrences and thematic networks revealed core themes such as classification, prediction, and model interpretation as central to the field, while also highlighting the growing role of neural networks and the challenges posed by black-box models. The thematic analysis further explains the structure of the field, distinguishing between foundational themes that anchor much of the research and more dynamic areas that drive innovation and advance the state of the art. In particular, the identification of niche and emerging themes suggests that XAI is not only expanding but also diversifying, with applications ranging from health diagnostics to risk assessment. The next step will be to refine and extend this analysis to gain a deeper understanding of the field and identify more specific avenues for future research. In addition, exploring the intersection of XAI with other fields, as well as the societal and ethical implications of these technologies, could provide fruitful avenues for further research.

References

Abusitta, A., Li, M. Q., Fung, B. C. (2024). Survey on explainable AI: Techniques, challenges and open issues. *Expert Systems with Applications*, 255, 124710.

Adadi, A., Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, pp. 52138–52160.

Aria, M., Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), pp.959–975.

Aria, M., Cuccurullo, C., Gnasso, A. (2021). A comparison among interpretative proposals for random forests. *Machine Learning with Applications*, 6, 100094.

Aria, M., Gnasso, A., Iorio, C., Pandolfo, G. (2024). Explainable ensemble trees. *Computational Statistics,* 39(1), pp. 3–19.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, pp. 245–317.

Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), pp. 203–216.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, pp. 285–296.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys* (CSUR), 51(5), pp. 1–42.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1), pp. 7–15.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., PRISMA Group, T. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine,* 151(4), pp. 264–269.

Molnar, C., Casalicchio, G., Bischl, B. (2020). Interpretable machine learning – A brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 417–431). Springer.

Whig, P., Kouser, S., Bhatia, A. B., Nadikattu, R. R., Sharma, P. (2023). Explainable machine learning in healthcare. In *Explainable Machine Learning for Multimedia Based Healthcare Applications* (pp. 77–98). Springer.

doi.org/10.26398/asaproc.00113

Innovative ER solutions: Using data to drive operational success

Maria Gabriella Grassia [a], Dario Sacco [a], Simone Paesano [a], Federica Dagostino [b], Angela Annachiarico [b], Valentina Di Palma [b]
[a] Department of Social Science University of Naples Federico II, Naples, Italy.
[b] A.O.R.N. Caserta, Napoli, Italy.

1. Introduction

In emergency room (ER) settings, the intricate relationship between patient wait times and the accuracy of triage codes is pivotal for ensuring effective patient care and outcomes (Hinson et al., 2020; Çetin et al., 2020). Wait times denote the interval from when patients arrive at the ER to when they receive medical evaluation or treatment, whereas triage codes categorize patients based on the urgency of their conditions, guiding priority in treatment. The dynamic between wait times and triage accuracy critically affects both the quality of care and patient satisfaction levels (Göktüğ et al., 2020). The research aims to provide a useful tool that can help health professionals now of triage attribution. This paper explores this relationship using data from AORN Caserta Hospital's emergency department collected from January 1, 2023, to December 31, 2023, comprising 67.676 observations. The variables considered include arrival and take-in times, color triage inputs and outputs, and diagnosis ICD-9 codes.

2. Literature review

The assessment of the triage process in emergency rooms is essential for patient safety and efficient healthcare delivery. Factors influencing patient safety during triage include accuracy, preparedness of staff, and the effectiveness of triage systems (Fekonja et al., 2023; Yuvita et al., 2022). Training programs for nurses can significantly impact their knowledge and practices related to triage, thereby improving patient outcomes (Hassan and Hussein, 2019; Mostafa et al., 2023). Implementing structured triage systems, such as the Emergency Severity Index (ESI) or the JTAS, can enhance the efficiency of emergency departments and reduce adverse events (Fajarini, 2022; Koyama et al., 2017). Collaborative approaches between triage nurses and clinicians, as observed in the context of COVID-19 triage, can streamline decision-making processes and enhance patient care (Poggiali et al., 2021). Additionally, factors like nurse competence, stress levels, and clinical reasoning skills can influence the accuracy of triage assessments (Hwang and Shin, 2022). In conclusion, a well-functioning triage process is crucial for timely and effective patient care in emergency settings.

3. Data and methodology

Given the complex dynamics of the emergency department described in the relevant literature, we formulated two research questions:

RQ1. Is there a similarity in the diagnosis attributed to ICD-9 coding across different triage color classifications?

RQ2. Are there diagnoses that systematically contribute to inappropriate wait times and triage categorizations during the discharge process?

To answer these questions, we performed an initial exploratory analysis, focusing on the adequacy of waiting times for triage codes. This analysis provided insights into the internal dynamics of the emergency department (Cremonesi et al, 2015). Specifically, we observed in *Figure 1* that 67.85% of orange-coded patients did not meet the guidelines for timely admission, highlighting potential inefficiencies in the triage process. Furthermore, 42.83% of these patients experienced changes in their triage codes between their initial evaluation (Triage In) and their final assessment (Triage Out).

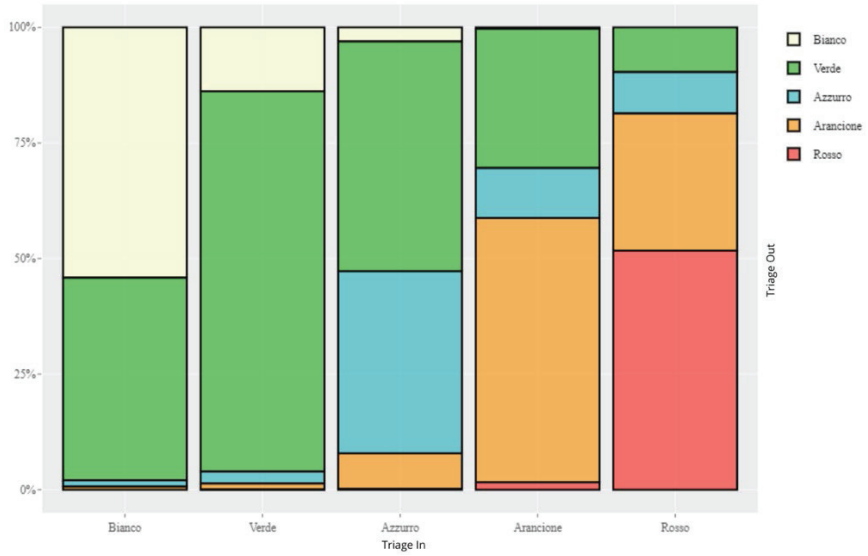


Figure 1: Triage code appropriateness

The data considered concern the records made in the ER of the AORN Hospital Company of Caserta during the period from 01/01/2023 to 12/31/2023. A total of 67.676 observations. The variables considered were: Time and hour of ARRIVAL; Time and hour of admission (TAKE IN); Color Triangle Intake; Color Triangle Exit; and Diagnosis ICD9 code.

The process begins with data from the "EMUR Flow," containing patient records with details on diagnosis and triage. The first step involves data preparation and cleaning. Relevant variables like diagnosis codes, triage categories, and timestamps are selected, then formatted and standardized. ICD-9 codes are pre-processed for consistency, followed by cleaning to remove incomplete or incorrect data and normalization to standardize text formats. Unnecessary words are filtered out, leaving key terms related to diagnoses and triage. The data is then divided into 10 subsets based on triage code and admission time.

Next, metrics are calculated. Diagnoses are compared pairwise using the Jaro-Winkler algorithm, creating 20 similarity matrices. Triage codes are similarly paired to track changes from admission to discharge, generating another 20 matrices. These analyses define two metrics: the Matching Index (MI) for diagnosis similarity (X-axis) and the Triage Change Index (TC) for triage variability (Y-axis).

Finally, visualization is achieved through the CARE Plot, which displays MI and TC to uncover patterns in diagnosis similarity and triage changes, providing insights into their alignment over time.

The appropriateness of waiting times was evaluated based on guidelines provided by the Ministry of Health. The Ministry establishes for each triage colour a time limit for taking charge of the patient, by which the assigned code becomes inappropriate. We coded this information as a

dummy variable.

Because 67.85% of patients with code orange do not meet ministerial guidelines on waiting times, we focused the analysis on these cases. Given that, 42.83% of these patients are given a different triage code on entry and exit, and that for 30.15% the orange code on entry becomes a green code on exit, we wondered about the possibility of common diagnoses and situations between unfit orange codes and fit greens. This might indicate that orange codes on entry could be classified as green.

Therefore, we considered two subsets for analysis: the first consisting of units with ineligible orange codes for waiting times, and the second consisting of units with eligible green codes for waiting times.

To address our first research question, we developed the Matching Index using the Jaro-Winkler Similarity metric to measure the distance between diagnoses having the two different triage classifications considered (Dumitrescu and Popescu, 2014; Rahmani et al, 2018).We adopted a threshold between 0.8 and 1, as, according to the results of the literature, this value is adequate when it is desired to ensure a common prefix of at least two letters and a correspondence of more than six among the remaining characters.

An initial matrix was defined, with rows representing diagnoses of non-eligible orange cases and columns representing eligible green cases.

The Matching Index (MI) is defined as the ratio of the i-th diagnosis count for the ineligible blue triage code to the total diagnosis count, for diagnoses exceeding the threshold of 0.8 according to the JW. This index was normalized to a scale ranging from -1 to 1 to facilitate uniform evaluation.

$$M_i = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} f_i}$$

$where$:
$M_i\ is\ the\ Matching\ Index$
$c_i\ is\ the\ i-th\ modality\ of\ the\ specific\ diagnosis\ of\ the\ non-eligible\ orange\ triage\ code,$
$which\ exceeded\ the\ threshold\ of$ 0.85 $measured\ by\ the$ Jaro − Winkler Similarity;

$f_i\ is\ the\ frequency\ of\ the\ i-th\ modality\ related\ to\ the\ specific\ diagnosis\ of\ the\ subset$

$$M_{norm} = \frac{[(M_i - \min(M_i)) * 2]}{Range(M_i)} - 1$$

To address our second research question, we developed the Triage Change Index to capture discrepancies in patient flow through the triage system. This index considered variations between initial and final triage assessments and was normalized to the same scale as the matching index. The index is normalized, as before, to allow for a uniform evaluation scale ranging from -1 to 1.

$$TC_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} f_i} \qquad \begin{cases} x_i = 1 \ \ if\ Triage_{in} \neq Triage_{out} \\ x_i = 0 \ \ otherwhise \end{cases}$$

$where$:
$TC_i\ is\ the$ Triage Change Index
$x_i\ is\ the\ i-th\ modality\ of\ the\ specific\ diagnosis\ of\ thesubset$;

$f_i\ is\ the\ frequency\ of\ the\ i-th\ modality\ related\ to\ the\ specific\ diagnosis\ of\ the\ subset$

5. Results

The indices revealed several key trends when plotted on a Cartesian plane shown in *Figure 2*. First quadrant: High Matching and Triage Change Indices. Diagnoses in this quadrant, despite frequent category changes, were common among eligible cases, suggesting a tendency to overestimate patient conditions. This finding indicates that some patients might receive more intensive care than necessary, potentially straining resources.

In this quadrant, we observe frequent triage classification changes and simultaneous common diagnoses between eligible and non-eligible cases, suggesting a tendency to overestimate patients' conditions. This finding indicates that some patients may be receiving more intensive care than necessary, potentially overstretching resources. For this classification, with a high probability, the orange triage assignment would be inappropriate.

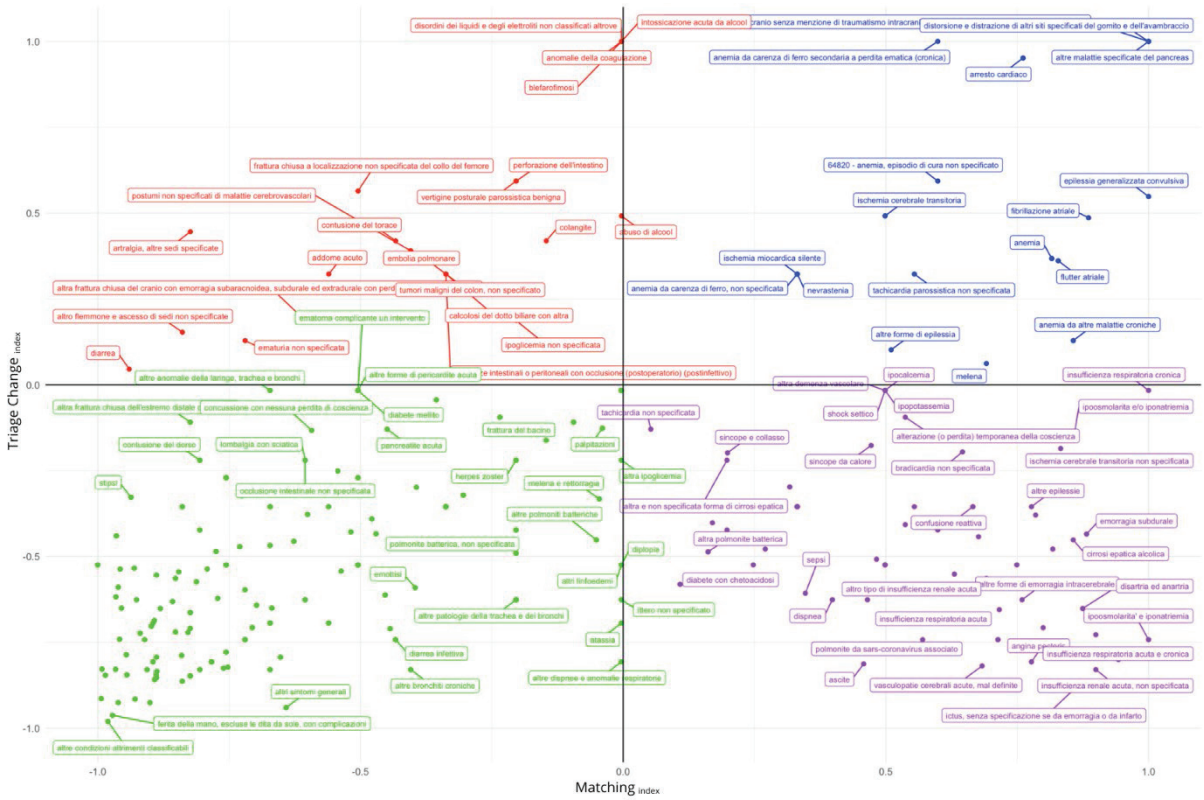


Figure 2 Accountability of triage

Second quadrant: Low Matching but High Triage Change Index. This area indicated frequent complications in cases, highlighting inefficiencies in the triage process that require improvement. Patients in this quadrant often experienced changes in their triage status, suggesting that their initial assessments might not have accurately reflected the urgency of their conditions.

Third quadrant: Low Matching and Triage Change Indices. This quadrant suggested that external factors, beyond the scope of this study, might be influencing the discrepancies encountered in patient evaluations. These external factors could include systemic issues such as staffing shortages, equipment availability, or broader healthcare policies.

Fourth quadrant: High Matching but Low Triage Change Index. This quadrant served as an alert indicator, where diagnoses were quickly classified as appropriate without changes from initial to final triage. This suggests that initial triage assessments were accurate, but the lack of change could also indicate hasty evaluations that did not fully capture patient needs.

6. Discussion and conclusions

Taking RQ1 and RQ2 again, we found numerous diagnoses that are classified with different colour-triage, high Matching Index values describe these cases. We find 35.12% of the ineligible orange codes with a Matching Index value above the threshold. Of these cases, the first quadrant in Figure 1 shows the diagnoses that also have high values of the Triage Change Index. So, in addition to not meeting ministerial waiting times and having many similarities between different colors, they change triage colour from entry to exit. These cases are 17.45%.

To further illustrate overall the application of our indices, we conducted detailed case studies on specific diagnoses that fell into each quadrant. For instance, in the Low Matching and High Triage Change Index quadrant, we examined cases of chest pain that were initially categorized as non-urgent but later required urgent intervention. This case highlighted the need for improved initial assessments to prevent delays in critical care (Chan, et al, 2005).

In the High Matching and Low Triage Change Index quadrant, we studied instances of minor injuries that were consistently categorized correctly, indicating effective triage processes. However, the high matching index also suggested that these cases might benefit from streamlined processes to reduce waiting times further. The findings from our indices and case studies underscore the importance of refining triage procedures to enhance both efficiency and effectiveness. The quadrant analysis provides valuable insights for improving patient management and adhering to health ministry guidelines regarding wait times. Specifically, addressing the issues highlighted by the indices can lead to better resource allocation and more accurate patient assessments

The quadrant characterized by low values in both indices emphasizes diagnoses that did not have serious problems with matching or triage variation during the emergency department process. This quadrant represents the best, in our case study, patient management.

The presence of overestimated conditions in the high matching and triage change index quadrant indicates a need for stricter evaluation criteria to prevent unnecessary admissions. This finding suggests that certain diagnoses may be over-prioritized, leading to resource inefficiencies. Conversely, the high matching but low triage change index quadrant highlights the potential for overly rapid assessments, which could compromise patient care quality. This suggests a need for a balanced approach that ensures thorough initial evaluations without unnecessary delays.

Our study enhances the understanding of triage procedures and identifies areas for further investigation and optimization. By analyzing the quadrants defined by the indices, healthcare providers can target specific inefficiencies in the triage process, ultimately improving patient outcomes. Future research should focus on the degree of change within the triage change index to further refine the evaluation system.

Future research should delve deeper into the external factors influencing triage accuracy and waiting times (Çetin et al., 2020). Additionally, longitudinal studies could provide insights into the long-term impacts of refined triage processes on patient outcomes. Another area of interest is the application of machine learning algorithms to predict and improve triage outcomes based on historical data.

References

Çetin, S., Eray, O., Cebeci, F., Coşkun, M., Gözkaya, M. (2020). Factors affecting the accuracy of nurse triage in tertiary care emergency departments. *Turkish Journal of Emergency Medicine*.

Chan, T. C., Killeen, J. P., Kelly, D., Guss, D. A. (2005). Impact of rapid entry and accelerated care at triage on reducing emergency department patient wait times, lengths of stay, and rate of left without being seen. *Annals of Emergency Medicine*, 46(6), pp. 491-497.

Cremonesi, P., di Bella, E., Montefiori, M., Persico, L. (2015). The robustness and effectiveness of the triage system at times of overcrowding and the extra costs due to inappropriate use of emergency departments. *Applied Health Economics and Health Policy*, 13, pp. 507-514.

Dumitrescu, S. R., Popescu, D. (2014). String identifier in multiple medical databases. *International*

Journal of Advanced Networking and Applications, 6(3), 2334.
Fajarini, M. (2022). The emergency severity index (ESI) usage: triage accuracy and causes of mistriage. *Jurnal Aisyah Jurnal Ilmu Kesehatan*, 7(S1). https://doi.org/10.30604/jika.v7is1.1190
Fekonja, Z., Kmetec, S., Fekonja, U., Reljić, N., Pajnkihar, M., Strnad, M. (2023). Factors contributing to patient safety during triage process in the emergency department: a systematic review. *Journal of Clinical Nursing,* 32(17-18), pp. 5461-5477. https://doi.org/10.1111/jocn.16622
Göktüğ, A., İlarslan, N., Vatansever, G., Özdemir, İ., Polat, O., Oğuz, A., … Tekin, D. (2022). Evaluation of the validity and reliability of ankutriage, a new decision support system in pediatric emergency triage. *Pediatric Emergency Care*.
Hassan, H. and Hussein, H. (2019). Effectiveness of education program in nurses' practices about triage system in emergency department at Galat Salih hospital. *Kufa Journal for Nursing Sciences*, 9(1). https://doi.org/10.36321/kjns.vi20191.2258
Hinson, J. S., Martinez, D. A., Cabral, S., George, K., Whalen, M., Hansoti, B., Levin, S. (2019). Triage performance in emergency medicine: A systematic review. *Annals of Emergency Medicine*.
Hwang, S., Shin, S. (2022). Factors affecting triage competence among emergency room nurses: a cross-sectional study. *Journal of Clinical Nursing*, 32(13-14), pp. 3589-3598. https://doi.org/10.1111/jocn.16441
Koyama, T., Kashima, T., Yamamoto, M., Ouchi, K., Kotoku, T., Mizuno, Y. (2017). A study of the effect of introduction of jtas in the emergency room. *Acute Medicine & Surgery*, 4(3), pp. 262-270. https://doi.org/10.1002/ams2.266
Mostafa, S., Emam, E., Youness, E. (2023). Training program impact on nurses' knowledge and practice regarding obstetric triage. *International Journal of Obstetrics and Gynaecological Nursing*, 5(1), pp. 56-62. https://doi.org/10.33545/26642298.2023.v5.i1a.108
Poggiali, E., Bastoni, D., Ferrari, M., Moretto, D., Buttafava, F., Ramos, P., … Vercelli, A. (2021). COVID-19 or not COVID-19, that is the question. The role of screening criteria and point-of-care lung ultrasound in the triage decision-making process in an emergency department during the phase 2 of the COVID-19 Italian epidemic. *Emergency Care Journal*, 17(2). https://doi.org/10.4081/ecj.2021.9708
Rahmani, F., Sepehri Majd, P., Ebrahimi Bakhtavar, H., & Rahmani, F. (2018). Evaluating the accuracy of emergency nurses in correct triage using emergency severity index triage in Sina Hospital of Tabriz: A cross-sectional analysis. *Journal of Emergency Practice and Trauma*.
Yuvita, E., Fitriasari, N., Harjayanti, N. (2022). Relationship between knowledge and preparedness levels of ER staff on pre-triage system implementation. *Jurnal Kedokteran Brawijaya*. https://doi.org/10.21776/ub.jkb.2021.031.02.8s

doi.org/10.26398/asaproc.00114

Responsible AI/ML implementation in official statistics: A framework for decision-making

Gerarda Grippo , Alessandra Righi
ISTAT – Italian National Institute of Statistics, Rome, Italy

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are revolutionizing statistical production, offering tools to meet the demand for timely, detailed, and high-quality data while ensuring ethical and secure practices. ML enhances efficiency by complementing or replacing traditional methods, but its adoption poses challenges beyond methodology, requiring collaboration across organizational departments and expertise in data science, ethics, and privacy. This paper identifies best practices for leveraging AI/ML, emphasizing skills enhancement, and compares international and national initiatives to develop responsible frameworks and collaborative strategies for implementing algorithms in official statistics.

2. Responsible AI/ML techniques implementation in official statistics

Artificial intelligence holds significant potential for statistical organizations, enhancing the efficiency of statistical production by automating certain processes or assisting humans in their execution. Machine learning solutions are used in National Statistical Institutes (NSIs) for various purposes, including data collection, integration, processing, and dissemination. These solutions can automate processes previously performed primarily by humans and enable organizations to utilize new data sources. The use of big data often necessitates adopting ML techniques, as it involves processing large volumes of data and exploring different application areas such as coding and classification, editing and imputation, and image analysis.

In recent years, private companies have also produced and disseminated statistics leveraging AI and ML techniques, providing their services promptly and making them particularly attractive to various stakeholders and policymakers. To remain competitive, institutions producing official statistics must organize themselves to harness the potential offered by AI and ML models, improving the relevance and quality of the official statistics produced and disseminated.

In this transformative and evolutionary journey, institutions must ensure the ethical and safe use of AI/ML models addressing technical, social, legal, and ethical challenges, with explainability, fairness, and quality as key components to avoid biased outcomes or privacy violations.

Finally, every organization faces limited resources to exploit the technology's potential. This makes it essential to invest in human resources, training, and hiring personnel with specific skills in these sectors, to facilitate the safe and ethical adoption of these new technologies.

3. International and national initiatives in progress

The awareness of the importance of skills and the sharing of knowledge and experiences as a decisive factor in facilitating the adoption of new technologies has led some international organizations to create discussion groups and think tanks to develop reflections in this field. The Applying Data Science and Modern Methods Group of the UN-ECE High-Level Group for the Modernization of Official Statistics (HLG-MOS) has started working on an international framework to support the official statistics community in developing AI systems and ML algorithms that respect human rights and privacy, ensuring they are fair, transparent, explainable, robust, secure, and protected (UN-ECE, 2024).

The project of the HLG-MOS on machine learning, conducted in 2019 and 2020 (HLG-MOS, 2023), concluded that editing and imputation were valid use cases for machine learning in official statistics production. As the adoption of ML methods for editing by statistical institutes has encountered some difficulties, the data editing task team of the Applying Data Science and Modern Methods (ADSaMM), established in 2022, decided to examine the reasons preventing the adoption of these methods and suggest guidelines for overcoming them.

In 2021, the Economic Commission for Europe of the United Nations published a report (UN-ECE, 2021) presenting practical ML applications within statistical organizations and a quality framework that could help guide the choice of methods and key steps to facilitate the transition of ML models from the experimental stage to the production stage. The five dimensions of the quality framework consider accuracy, explainability, reproducibility, timeliness (ML algorithms lead to significant savings in processing time), and cost-effectiveness. The main areas of application for ML include coding and classification, editing and imputation, and image analysis. In coding and classification, studies showed that even organizations with basic ML expertise could achieve strong predictive results through collaboration and code sharing. For editing and imputation, ML demonstrated faster and more consistent performance compared to manual editing, while providing results comparable to traditional methods in a more automated manner. In image analysis, pilot studies reported model accuracy ranging from 74% to 97%, highlighting ML's iterative nature and its potential for continuous improvement with experience.

The International Monetary Fund (IMF) used machine learning for experimental and research applications, including GDP nowcasting, developing proxy measures of trade using ship positions, and coding economic time series through textual analysis. With the increasing availability of generative artificial intelligence tools, the IMF is also exploring how these techniques can improve its data services, from data collection to dissemination (Marini, 2023).

In 2024, Eurostat launched the ESSnet one stop-shop AI/ML for official statistics (*https://cros.ec.europa.eu/dashboard/aiml4os*) project, lasting 4 years, with the participation of 14 NSIs led by Ireland's Central Statistics Office, aimed at creating a sandpit platform/hub, developing capabilities, and building communities around solutions, providing support and guidance for the integration and maintenance of AI/ML solutions through training, communication, and dissemination of results. The ultimate goal is to enable the transition from experimental solutions to official statistical production solutions. Istat leads the activities relating to communication and community engagement, tasked with raising awareness around these topics and engaging the AI and ML expert community beyond the statistical community. Istat is also responsible for the use case generating synthetic data balancing utility and privacy.

4. A focus on Istat's experiences

In its first pilot study, conducted following UNECE methodological and ethical guidelines, Istat applied ML tools to improve editing processes in the Italian public administration register, using data from the Public Administrations Database (BDAP) and the Information System on the Operations of Public Bodies (SIOPE). Methods included identifying data inconsistencies and applying decision trees and random forests for classification. Results demonstrated the potential of ML to streamline editing and imputation processes, enhancing accuracy and efficiency (UN-ECE, 2021).

In the second pilot study on data imputation, Istat imputed the variable "Attained level of education" in the base register of individuals. Data included administrative information from the Ministry of Education, University and Research, 2011 census data, and sample survey data. The investigation aimed to determine the benefits of ML over classic statistical models for imputation issues. Focusing on one region and incomplete records, manual feature selection, and prediction comparisons were performed using algorithms like MLP, random forests, and log-linear model, with Python and Azure on a cloud platform using Tesla V100-PCIE-16GB GPU. Accuracy was

measured at both micro and macro levels. Results indicated that MLP and log-linear models yielded comparable estimation results, with log-linear imputation performing better for extreme items (e.g., PhDs), while MLP showed slightly better micro-level accuracy and did not require pre-treatment of variables.

The pilots followed ethical and methodological guidelines by prioritizing transparency and explainability through the use of clear ML algorithms like decision trees and random forests. Accuracy was a key focus, with metrics assessed at micro and macro levels for editing and imputation processes. Collaboration across sectors and shared platforms emphasized interdisciplinary knowledge sharing, while ethical considerations ensured proportional data use and respect for privacy in managing microdata.

Integrating these techniques into production processes requires a multifaceted approach that intersects and spans the various organizational structures of a statistical institute. This includes: i) the human resources sector for skills development and the acquisition of new resources; ii) the IT sector for infrastructure; iii) the legal sector for ethical considerations; iv) the international relations, fostering continuous dialogue among different experiences and contexts to find shared solutions on these specific topics (De Cubellis and Grippo, 2023; 2024).

Human resources ensure that employees' competencies align with the IT infrastructure and are updated according to evolving national and international statistical contexts. Istat organizes specific courses on this topic and promotes the participation of the staff in the European Master in Official Statistics programs and European Statistical Training Programme courses, as well as courses organized by the (Italian) National School of Public Administration and universities. Doctoral scholarship for these specific topics are forthcoming. Internships have also been activated to allow university students to conduct part of their thesis at the Institute on projects developed using ML techniques.

Partnerships with academic institutions and private sector companies enhance skills, share knowledge, and bring diverse perspectives. Fostering a collaborative culture internally is crucial, encouraging interdisciplinary work among statisticians, data scientists, IT professionals, and subject matter experts to ensure that ML solutions align with business needs and are integrated into statistical production processes. Istat has taken strategic steps to ensure this collaborative approach. In 2020, they established the Center for Trusted Smart Statistics, an interdepartmental body to monitor and coordinate ML projects. In 2018, they inaugurated an innovation laboratory, based on the Dutch Central Bureau of Statistics' experience, to promote research and innovation. Projects in the Innovation laboratory are proposed and developed by multidisciplinary teams, integrating diverse skills to improve statistical production (De Cubellis and Grippo, 2023; 2024).

Ethical considerations are vital in AI/ML, particularly in managing microdata and private data. Ensuring transparency and public trust requires proportional data use and ethical guidelines across all ML stages. Ethical challenges, present throughout statistical production, call for systematic solutions and cross-sector collaboration to promote transparency and validation through shared methodologies.

5. Conclusion

AI and ML offer opportunities and challenges for statistical institutes. European projects are advancing, and international frameworks by UNECE and Eurostat guide the ethical, efficient use of AI/ML. Addressing ethical and methodological issues and fostering collaboration will help meet the demand for reliable statistics, despite limited resources and technological constraints. Istat's experience underscores the importance of a multifaceted approach, involving various organizational sectors such as human resources, IT, legal, and international relations. Empirical results coming from this framework will only become measurable in the coming years, as it is an ongoing process that has yet to produce fully comparable outputs. Partnerships with academic institutions and private sector companies, along with fostering a collaborative culture, are essential for successful implementation.

References

De Cubellis, M., Grippo, G. (2023). Trusted smart statistics: New statistics for decision makers. The Istat's experience. In *Data-Driven Decision Making*, Firenze University Press and Genova University Press, pp. 125–128.

De Cubellis, M., Grippo, G. (2024). Organizational sustainability to support trusted smart statistics: Istat's experience. *Statistical Journal of the IAOS*, pp.1 – 14.

High-Level Group for the Modernisation of Official Statistics (HLG-MOS) (2023) *Large Language Models for Official Statistics*. HLG-MOS White Paper. Modernstats HLG-MOS.

Marini, M. (2023). Responsible AI for data and statistics at the IMF, *64th ISI World Statistics Congress - Ottawa, Canada*.

Moritz, S., Sirello, O. (2024). Organisational aspects of implementing ML based data editing in official statistics. *Conference of European Statisticians Expert Meeting on Statistical Data Collection and Sources 22-24 May 2024, Geneva, Switzerland*.

Stramer, K. (2020). *UNECE – HLG-MOS Machine Learning Project Classification and Coding Theme.*https://unece.org/sites/default/files/202201/ML_06_Classification%20and%20Coding%20Theme%20Report.pdf

UN-ECE (2021). *Machine Learning for Official Statistics*, Geneva, UN.

UN-ECE (2024). *Follow-up on the Conference of European Statisticians Seminar "Data Ethics – A Key Enabler of Social Acceptability*, *Geneva, 20 and 21 June 2024* Item 8, ECE/CES/2024/19.

doi.org/10.26398/asaproc.00115

Copula functions and dynamic hydrological data: The case of Scotland

Sara S. Guerini [a], Luigi Spezia [b]
[a] *Department of Economics, University of Bergamo, Italy.*
[b] *Biomathematics & Statistics Scotland, Aberdeen, UK.*

1. Introduction

Due to the global warming, the patterns of seasonal precipitation in Scotland are changing more and more each year. These changes could lead to increased periods of droughts, which consist of prolonged periods of low rainfall, that lead to water scarcity (Centre of Expertise for Waters, 2024). On the other hand, the intensity of extreme precipitation events and the risk of flooding are also increasing. To evaluate these problems, the Scottish Environment Protection Agency (SEPA) takes records of the discharges of a set of streams and rivers in Scotland. Discharges are recorded on an hourly basis using a variety of electronic sensors and data loggers. Then, these values are combined on a daily basis to monitor the flow dynamics.

The aim of this work is to investigate the suitability of dynamic copula models in hydrological applications, especially in detecting when periods of either water scarcity or water abundance occur. Hydrological processes exhibit serial dependence, making them challenging to model accurately. We investigate the ability of copula models to produce satisfying results in hydrology, with a focus on their effectiveness in capturing the complex relationships among hydrological variables and detecting extreme events. The primary objective of our analysis is the detection of significant variations in river or stream discharges. Thus, the development of statistical techniques to effectively investigate features of serially dependent data is needed.

2. The data set

The data set we consider is a time series of 20 years of daily measurements of the discharges (in m^3/s) recorded in the stream Luggie Water at the Condorrat gauging station (coordinates (273917,672559)) which is located between the cities of Glasgow and Edinburgh, the most populated area of Scotland.

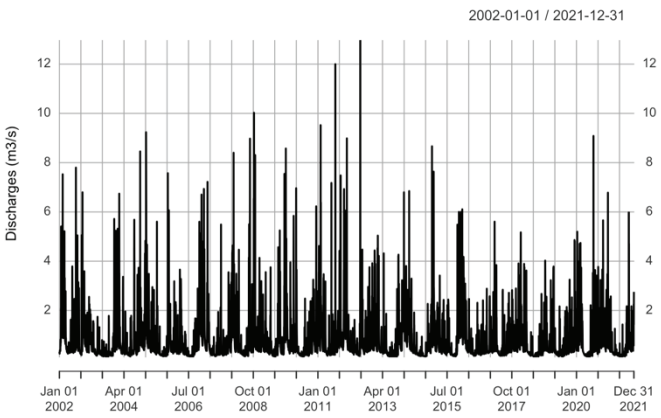


Figure 1: Time series of daily discharges in Luggie Water

The series (Figure 1) looks somewhat volatile. In fact, we can detect peaks in the values that may indicate abundance of water and time intervals in which the values are really low, suggesting water scarcity.

As expected, the highest and the lowest values in the discharges happen with a yearly periodicity. In fact, periods of flood are observed mainly in the autumn and the winter, whereas periods of water scarcity during the warmer months.

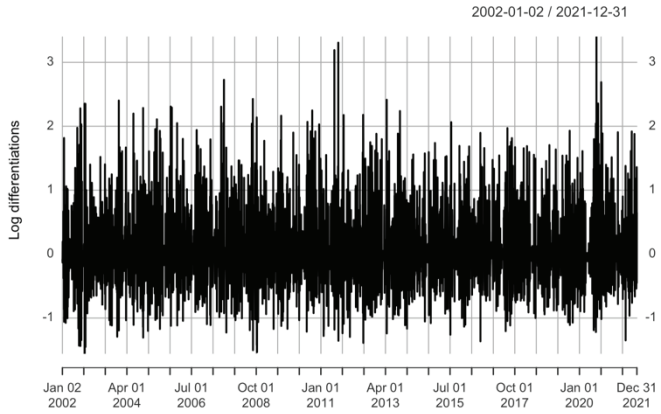


Figure 2: Log-differentiation of the series in Figure 1

To better apply copula models, we need to transform the original series $\{Y_t\}$ of Figure 1 into a new series $\{X_t\}$ (Figure 2), where $X_t = \log(Y_t) - \log(Y_{t-1}) = \log\left(\frac{Y_t}{Y_{t-1}}\right)$, for each value $t = 2, \dots, T$.

Further transformations, i.e. $\{U_t\}$ and $\{V_t\}$, are introduced and they are referred as volatility proxy transformations. Variables U_t are the standardized ranks of X_t, whereas V_t are the standardized ranks of the absolute value of X_t $(t = 2, \dots, T)$, where standardized rank means that the transformed variables take a value from 0 to 1 which will indicate its position in the ranking, with 0 being the day in which we have the lowest discharge, and 1 is the day in which we have the highest discharge. These transformations are called pseudo-copula, and the distribution of $\{V_t\}$ against $\{U_t\}$ is presented in Figure 3. These transformations allow modelling the variations between observations and highlighting their variability.

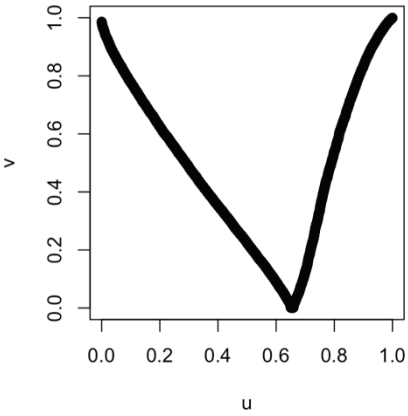


Figure 3: Distribution of {Ut} against {Vt}

3. The model

In order to carry out this analysis, the application of copula functions and V-transforms (where V stands for vine) is proposed. Copulas link (or couple) multivariate density functions to their univariate margins. Hence, they can be applied to study the dependence between the components of a random vector. Vine copulas can be seen as copula models of dependence. These are used to handle non-Gaussian marginal behaviours and non-linear and non-Gaussian serial dependence behaviours. More specifically, the method is used to define a class of tractable non-Gaussian autoregressive moving average (ARMA) processes and non-Gaussian AR fractionally integrated MA (ARFIMA) processes. The V-transforms describe the relationship between the quantiles of the stationary distribution of the time series and the quantiles of the distribution of a predictable volatility proxy variable. They can be represented as copulas, allowing formulation and estimation of models that combine arbitrary marginal distributions with copula processes for the dynamics of the volatility proxy.

The best configuration of vines for time series and longitudinal data is the D-vine (where D stands for drawable), which can describe the strict stationarity of a vector, following some translation-invariance principles on the vine structure.

VT-ARMA copula processes are stationary time series processes obtained by using a V-transform and an ARMA process.

4. Inference and model choice

In our modelling, we apply the semiparametric estimation method developed by Genest et al. (1995) and make use of the R package `tscopula` (McNeil and Bladt, 2024). This approach has

been employed by Chen and Fan (2006) for the first-order Markov copula models. First, we estimate the marginal distribution nonparametrically. This transformation is applied to re-scale the data to the (0, 1) interval. Then, a pseudo-copula data set is created by applying some transformations to the dataset, which in our case was the standardised rank transformation, and used to estimate the parameter of the copula model. This is done through the maximum likelihood estimation that finds each fulcrum, i.e. the maximum points of the likelihood functions. In our application, we initially use the semiparametric approach to find an effective copula process.

Then, the next steps consist in re-fitting the model with the estimates that we found, and, finally, finding the best models through the Akaike Information Criterion (AIC). Finally, we can compute the AIC value for each model and select the best one.

5. Results

The VT(3)-ARMA(1,1) model showed the lowest AIC value. However, all other models gave values of the same order of magnitude. Hence, further analyses were carried out on a set of best models. After some visual model diagnostics on these models as the QQ-plot and the autocorrelation function plot (Figure 4), we could conclude that these models are reasonable and could be used for further analyses. Thus, we could fit the copula models to our data, and afterwards add a parametric marginal distribution to see if the model fit improved. For all of the models, we fitted margins that follow the Normal, the Student-t, and the Laplace distributions. Each of these distributions has different properties, shapes, and characteristics, which can impact the modelling process and the interpretation of results. The Normal distribution is used when the data are approximately symmetric and unimodal. The Student-t distribution can capture heavy tails and more extreme values in nearly symmetric samples. Laplace distribution, or double exponential distribution, is symmetric with heavy tails, too, and is well robust against outliers.

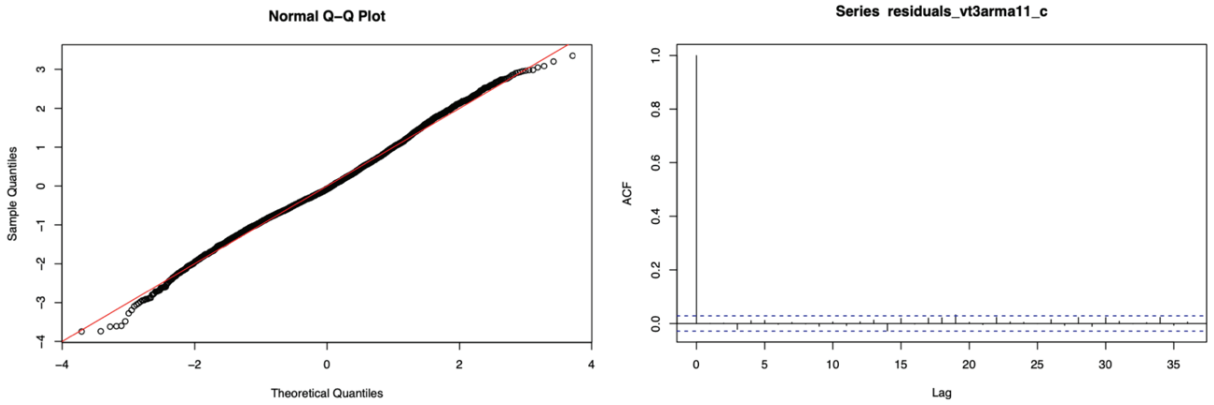


Figure 4: QQ plot and ACF plot of the residuals of the VT(3)-ARMA (1,1) model

For each model, first we created the margins, fitting the margins to the data. As a second step, the margins that were previously calculated needed to be applied to the model result. Afterwards, we could decide which margin was the best for each model by calculating updated AIC values.

Model	*Normal*	*Student-t*	*Laplace*
Vt3arma11	6358	4210	**4104**
Vt2arma11	4998	4133	**4100**
Vt1arma11	5477	**4322**	4459
Vt1arma40	6452	**4323**	4475
Vt1arma30	6651	**4340**	4497

Table 1: AIC values for each model, with specific margins. In bold the AIC associated with the best models for each row.

We compared five different models (Table 1) which differ either in the ARMA order or in the order of the V-transform. Looking at the results, the preferred margins are the Student-t margin and the Laplace margin. An interesting fact seems that the model that include both the autoregressive and the moving average seem to favour (even if only slightly) the Laplace margins, while the models with only the autoregressive part seem to favour the Student-t margin. VT(2)-ARMA(1,1) model is the best model

according to the AIC analysis and the best margins associated with this model are the Laplace margins.

Looking at the combinations of the VT(2)-ARMA(1,1) model with Student-t margins (Figure 5, left) and Laplace margins (Figure 5, right), a difference is evident, When the Student-t is used, the result is quite good for the negative values, which means that the model is good in modelling the days that have a decrease with respect to the previous day. On the other hand, when the Laplace is used, a better fit is evident. In fact, we see that there is a correct representation of the middle part of the data, but there are also some peaks and some drops that follow the data dynamic behaviour.

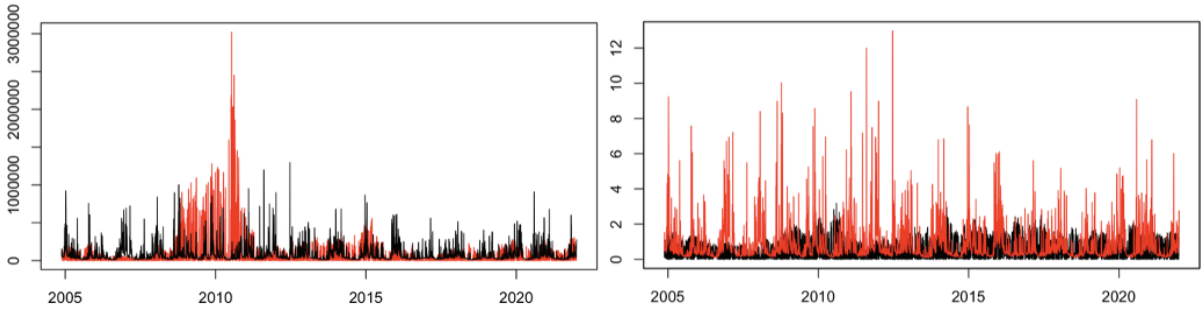


Figure 5: Out-of-sample forecasting for VT(2)-ARMA(1,1) model with Laplace (left) and Student-t (right) margins, in red, against the $\{X_t\}$ values in black.

By plotting the predictions resulting from the combination of VT(3)-ARMA(1,1) model and the Laplace margin (Figure 6, left), the values are concentrated around the mean, with no evident peaks, but with some representation of droughts, that seem to follow the drops in the X_t series. We see that the model follows the mean values quite nicely, but the peaks are not represented. Looking at the combination of the VT(3)-ARMA(1,1) model with Student-t margins (Figure 6, right), the representation seems to be similar to the Laplace margins, which indicates that for this model there is no strong preference between the Laplace and the Student-t margins.

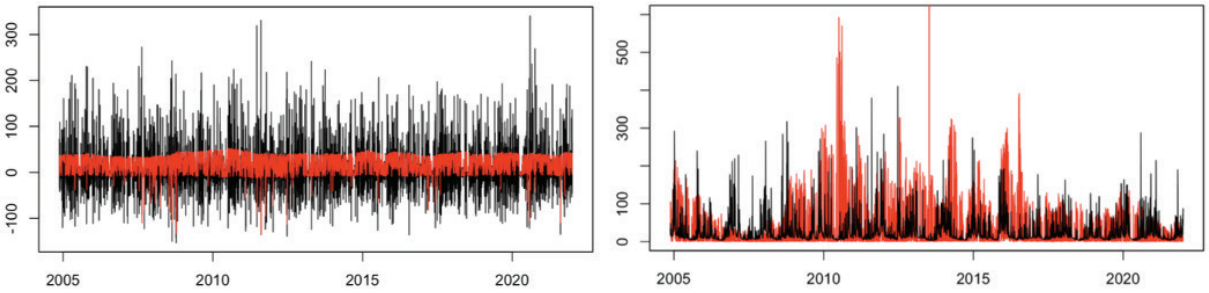


Figure 6: Out-of-sample forecasting for VT(3)-ARMA(1,1) model with Laplace (left) and Student-t (right) margins, in red, against the $\{X_t\}$ values in black.

The combinations coming from the VT(1)-ARMA(1,1) model seem to behave in the opposite way with respect to the other models that consider both the autoregressive and the moving average parts of the model. In fact, in this case, the best margin is the Laplace margin, while for the other ARMA models, the best combination was with the Student-t margin. If we look at the plot of the model together with the Laplace margins, shown in Figure 7 (left), the behaviour is good for the positive values, but it is absent for the negative values. This means that the model considers the positive day-by-day changes, but not the negatives. Moreover, if we see the plot for the combination with the Student-t margin, showed in Figure 7 (right), the fitting is not ideal, since it does not seem to follow peaks or drops, but it is almost a straight horizontal line apart a few peaks.

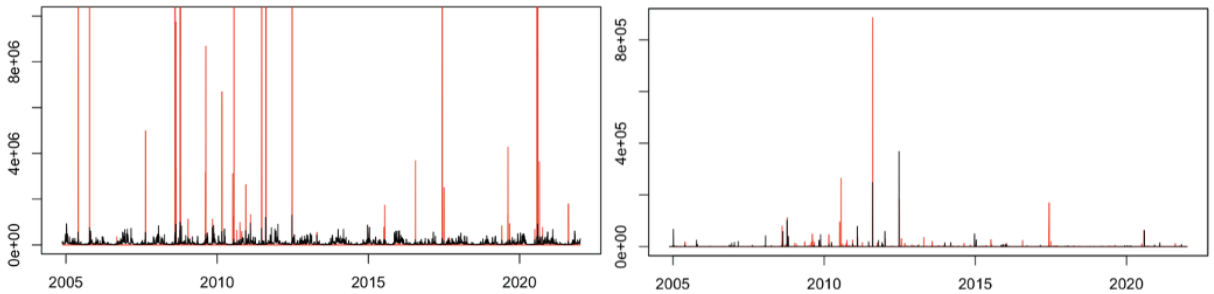


Figure 5: Out-of-sample forecasting for VT(1)-ARMA(1,1) model with Laplace (left) and Student-t (right) margins, in red, against the $\{X_t\}$ values in black.

Next, we consider the VT(1)-ARMA(4,0) model. First, we look at the X_t values against the results of the model with Student-t margins in Figure 8 (left). The fitting seems to be quite good for the negative values, meaning that the fitting of the negative differences with respect to the previous day look satisfactory. Different from the Student-t margins, the combination with the Laplace margins seen in Figure 8 (right) looks far from ideal. In fact, the modelling is a straight line with a very few drops. Hence, we can see that even if the AIC values are quite close together, the fitting is very different.

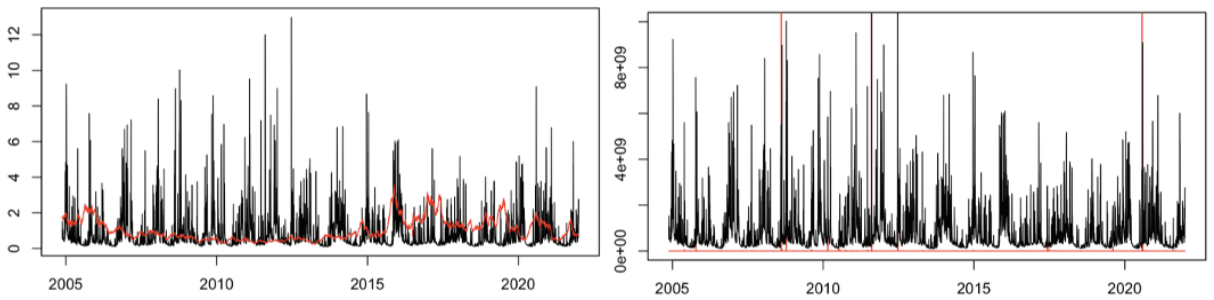


Figure 6: Out-of-sample forecasting for VT(1)-ARMA(4,0) model with Student-t (left) and Laplace (right) margins, in red, against the $\{X_t\}$ values in black.

Lastly, we look at the VT(1)-ARMA(3,0) model. Also in this case, the best combination is that with the Student-t margin and the Laplace is a close second. As we saw for the VT(1)-ARMA(4,0) model, if we represent the model with the margins together with the X_t data, there is a great discrepancy. In fact, in the plot regarding the Student-t margins (Figure 9, left), the fitting seems correct, even if only for the negative factors, while for the plot regarding the Laplace margins (Figure 9, right), the fitting is a flat line with only few drops.

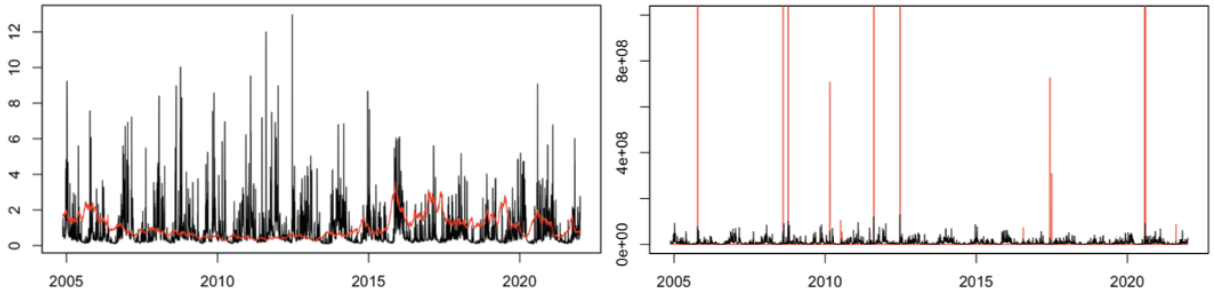


Figure 7: Out-of-sample forecasting for VT(1)-ARMA(3,0) model with Student-t (left) and Laplace (right) margins, in red, against the $\{X_t\}$ values in black.

Overall, the combination of model and margins that seems to be the best one is the VT(2)-ARMA(1,1) model with Laplace margins.

6. Conclusions

The results in the applications were also different among each other, and they varied according to which of the models we studied was employed. For example, we were able to infer that there are some combinations of models and margins (VT(2)-ARMA(1,1) model with Laplace margins) that capture the data better than other combinations (VT(1)- ARMA(4,0) model with Laplace margins). We also found that different models worked better with certain margins, whereas others worked better with other margins. We can say that applying dynamic copula models to serial dependent data, more specifically in the hydrological sector, is a promising and deserve further investigation in combination with the extreme value theory. In fact, we saw that there are good results coming from the modelling of the data. We have models that could advice in detecting extreme behaviours of the discharges, when the water levels are low or high.

Acknowledgements

SSG's research was funded by the European Union – Next Generation EU program via the PRIN PNRR 2022 project "SIGNUM" (CUP: F53D23010910001); LS's research was funded by the Scottish Government's Rural and Environment Science and Analytical Services Division. The data set was provided by the Scottish Environment Protection Agency (SEPA).

References

Chen X., Fan Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, **130**, pp. 567–577.

Centre of Expertise for Waters (2024). *Future Predictions of Water Scarcity in Scotland: Impacts to Distilleries and Agricultural Abstractors*. CREW Report, https://www.crew.ac.uk/publication/water-scarcity-impacts-distilleries-agricultural

Genest C., Kilani G., Rivest L.P. (1995). A semi-parametric estimation procedure of dependence

doi.org/10.26398/asaproc.00116

AI-generated sentiment in state-space models for Toyota Camry US volume sales forecasting

Domenica Fioredistella Iezzi[a], Roberto Monte[b], Daniele Pasquini[a]

[a] Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome, Italy

[b] Department of Civil Engineering and Computer Science Engineering, Tor Vergata University, Rome, Italy

1. Introduction

In recent years, generative artificial intelligence (AI) has been the focus of extensive attention from various fields of research (Mariani *et al.*, 2022). Generative AI is an artificial intelligence system that can create content based on human input. These systems use complex algorithms and machine learning techniques to learn patterns from existing data and generate new content accordingly. Examples of generative AI include systems that can produce language, images, videos, and audio (Giordano *et al.*, 2024). Large Language Models (LLMs) and generative AI can become valuable tools in sentiment analysis, particularly for disambiguating complex sentences. They are instrumental in examining how written words convey people's attitudes, sentiments, and emotions (Assiri et al., 2024). According to Farhat et al., 2023, LLMs like GPT-4 can generate contextually relevant, coherent, and complex responses, moving beyond essential keyword recognition to a more holistic understanding of human language.

In previous papers, we experimented with building a proxy of consumers' sentiment toward consumer goods to predict their sales by analyzing related communication in social media with ruled-based and machine-learning approaches (Iezzi and Monte, 2024a, 2024b). In this paper, we aim to build the proxy of consumers' sentiment toward the automobile Toyota Camry sold in the US by applying ChatGPT on a set of tweets and incorporating it as an observed signal in a state-space model for sales forecasting. ChatGPT is a proprietary service offered exclusively through platforms managed by OpenAI and does not qualify as free or open-source software. It is an AI-LLM designed to generate human-like text, providing responses that simulate natural conversation. Capable of addressing a wide array of topics, ChatGPT can answer questions, engage in dialogue, and create content across diverse domains. Currently, various versions of ChatGPT are available. First is ChatGPT-3.5, an open-access version released in November 2022, trained with 175 billion parameters up to September 2021. Second is ChatGPT-4.0, a premium version released in March 2023, trained with 100 trillion parameters up to September 2021. Studies show that ChatGPT-4.0 performs more precisely than ChatGPT-3.5, demonstrating greater intellectual capacity and improved skills in answering academic questions (Eysenbach, 2023).

To forecast sales volumes, we use different models and compare the results: an ETS-AAA model, with Additive error, Additive slope component in the local trend, and Additive seasonality; an ETS-ANA Model, with an Additive error, No-additive slope component in the local trend, and Additive seasonality; two Multivariate Autoregressive State-Space (MARSS) models distinguished by univariate and multivariate observations (Hyndman et al., 2002; Harvey, 1990; Holmes et al., 2020). To implement the MARSS models with multivariate observations for exploiting consumers' feelings, we have collected a Twitter corpus of 163,000 tweets. This paper is structured as follows: Section 2 presents ChatGPT 4.0 for sentiment analysis; Section 3 presents the main results of our models; Section 4 illustrates the main results of our case study.

2. ChatGPT 4.0 for sentiment analysis

The Generative Pre-Trained Transformer (GPT) model, best known as the engine behind ChatGPT, holds a prominent position in AI management, unlike the VADER (Hutto & Gilbert, 2014), which is a rule-based sentiment analysis tool specifically designed for analyzing social media texts and the BERT (Devlin et al., 2018), which uses bidirectional encoding, ChatGPT employs unidirectional encoding, meaning it which relies solely on the preceding words to understand context during text generation. This absence of masked training during pre-training marks a key difference from BERT. ChatGPT relies on the transformer architecture, featuring components like input embeddings, transformer encoders, self-attention mechanisms, and feedforward neural networks. It has shown remarkable versatility in handling zero-shot, one-shot, and few-shot tasks, effectively leveraging its pre-existing knowledge for new tasks without fine-tuning (Hwang Kwon et al., 2024). ChatGPT has garnered significant attention since its release, becoming a focal point of public discourse across various communities. In January 2023, merely two months after its launch, it reached an estimated 100 million monthly active users. It has sparked people's imagination about a revolutionized future in programming, writing, and many other fields (Koonchanok *et al*., 2023). ChatGPT has demonstrated its ability to enhance human productivity, shifting human work from the laborious tasks of code and text generation to problem formulation and higher-level thinking. Both GPT-4 and GPT-3.5-turbo are OpenAI's proprietary models. Unfortunately, until this publication, there is no confirmed public information regarding the training datasets or the pipelines OpenAI used to prepare these models (OpenAI, 2023; Ali & Onur, 2024). We have utilized OpenAI's paid API to fine-tune our models with our datasets and to gather responses for our prompts. We have chosen to work with GPT because, according to various studies, ChatGPT demonstrates superior linguistic accuracy compared to LlaMA and Mistral (Hou, 2024). It improves coherence and relevance across linguistic tasks (Bogireddy, 2024).

Additionally, in sentiment analysis tasks, ChatGPT performs well in open-domain scenarios and exhibits more robust capabilities in sentiment classification than BERT and state-of-the-art models (Wang, 2023). ChatGPT can process large amounts of text, identifying users' emotions, opinions, and attitudes. The advantages of using ChatGPT include its ability to recognize the word context and improve the accuracy of interpreting complex sentiments compared to keyword-based models. However, it may encounter difficulties with ambiguous or sarcastic texts where the actual sentiment is not immediately evident. Additionally, the quality of the analyses depends on the training data. Limited or low-quality data can influence the results.

In our analysis, we derive a sentiment proxy using 163,000 tweets about the Toyota Camry posted from June 2009 to December 2020. Table 1 shows some examples of sentiment classification of ChatGPT, using a score from -1 to +1, where -1 represents the most negative sentiment and +1 the most positive, values between -0.5 and +0.5 are neutral. A neutral sentiment indicates that the analyzed content does not express solid emotions or judgment-driven opinions. Therefore, the text is missing connotations when classifying it as either favorable or unfavorable. For example, informative or descriptive sentences like 'The car is red' or 'Today is Tuesday' are typically considered neutral because they do not convey judgments, emotions, or preferences.

Table 1 shows some examples of sentiment measured by ChatGPT for Toyota Camry tweets. The first tweet uses the phrase "100% APPROVAL," which is highly positive and suggests complete satisfaction or endorsement. The extensive use of exclamation marks amplifies the positive sentiment, indicating enthusiasm and excitement. The capital letters further emphasize a strong, positive emotional tone. The second tweet uses negative language because trading a Toyota Camry for another brand (Hyundai) implies dissatisfaction or disappointment toward Toyota.

Moreover, using the middle finger emoji directed at Toyota indicates strong negative emotions, such as anger or frustration. The exclamation mark emphasizes the negative sentiment, showing strong emotion. The third tweet is neutral since it is purely informational: stating that a 2003 Toyota Camry is for sale at Daytona Beach does not express emotions, opinions, or judgments. We can note no words or symbols that convey positive or negative sentiments. It is a straightforward, neutral statement. In the last tweet, we can see slightly positive language. It is a promotional tweet inviting potential buyers to consider a 2012 Toyota Camry. The phrase "In the market for" is neutral, but the context suggests a positive opportunity for those interested in purchasing a car. Also, in this tweet, as in the third tweet previously showed, we can notice the absence of Strong Emotional Cues**.** The tweet is prevalently informative but has a slightly positive connotation to attract consumers' interest.

Table 1 *Sample of Sentiment Classification on Toyota Camry by ChatGPT*

Text	**Sentiment score**
TOYOTA CAMRY 100% APPROVAL !!!!!!!!!!!!!!!!!!!!!!!!!!	+1
@axios Trading in my @Toyota Camry for a @Hyundai this afternoon. 🖕 U @Toyota!	-0.9
2003 Toyota Camry for sale - Daytona Beach	0
In the market for a #Toyota Camry? We've got a 2012 Toyota #Camry L 4dr Sedan	+0.4

As we can observe from the sample in Table 1, the sentiment analysis performed by ChatGPT considers at least five essential elements:

1. **Sentiment language:** negative, positive, or neutral words. This aspect is crucial in both dictionary-based sentiment analysis approaches and those based on machine learning and deep learning.
2. **Presence of emojis and emoticons:** to provide a more precise connotation of the direction and intensity of the sentiment.
3. **Use of punctuation:** to allow for the emphasis of sentiment expression, amplifying its impact.
4. **Capitalization:** capital letters strengthen the emotional tone, often associated with intense positive or negative feelings.
5. **Links:** in the presence of active links in the tweet, ChatGPT can learn its contents and use them to measure sentiment.

The sentiment of language is ubiquitous in all sentiment quantification, but ChatGPT also considers other aspects that different approaches seldom treat. Hence, sentiment quantification with ChatGPT certainly has the possibility of working using a multidimensional approach. As a language model, the advantage of ChatGPT is that it includes millions of documents and information from different sources, but we cannot check the quality of the data.

3. Main results

The sentiment analysis of Toyota shows a marked negative skew. Negative skewness highlights the presence of a pronounced left tail, signifying negative outlier values, alongside a higher concentration of data on the positive side. Notably, even a few negative reviews can significantly impact consumer behavior. Varga and Albuquerque (2024) demonstrate that when consumers encounter negative reviews, they are more likely to stop searching for additional reviews about the product. This reaction suggests a reluctance to invest further time and effort in gathering more information about that specific item.

In Figure 1 (a), the histogram highlights a long tail on the left side of the distribution. In other words, most of the data accumulates on the right side of the chart, while the left tail is

more extended, indicating a strong presence of tweets with positive sentiment. Notably, most tweets have neutral sentiment scores (ranging from -0.3 to +0.3). We also used a violin plot to visualize the sentiment distribution (Figure 1 (b)), combining the features of a box plot and a kernel density plot. Aside from the asymmetry, we can observe the density: the broader parts of the violin represent areas where the data is denser (thus on the right, corresponding to neutral and positive values), while the narrower parts indicate areas with fewer data points (on the left, where we find negative values).

Figure 1 Distribution Sentiment of tweets calculated with ChatGPT – Histogram and Violin Plot

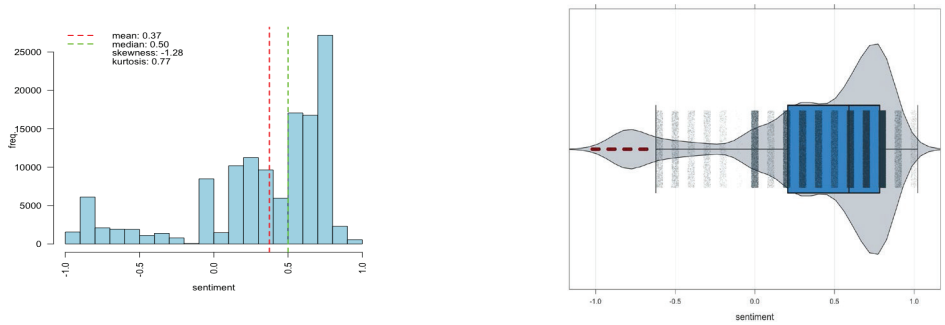
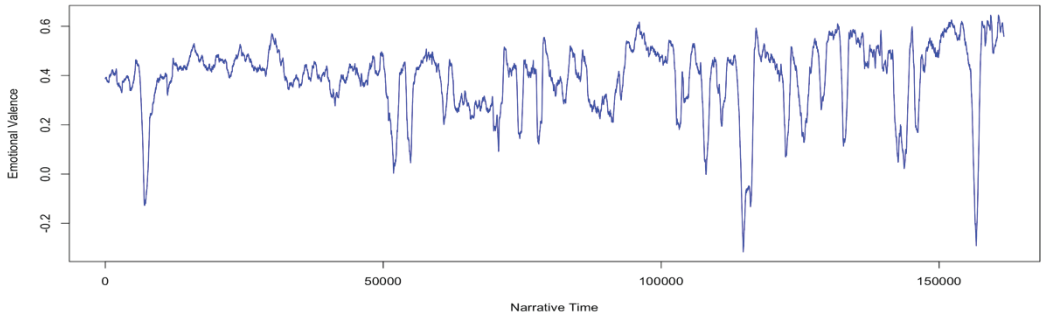


Figure 2 shows the plot of 'emotional narrative time,' which represents how a narrative is shaped or perceived in time in terms of emotions about both characters and the audience. To generate Figure 2, we utilized our tweet time series (approximately 163,000 tweets) and applied a moving average with a window size of k=1000 to illustrate the trend. We can see that the most significant negative peaks correspond to reports of thefts, issues with airbags, minor technical problems, and the COVID period.

Figure 2 Emotional Narrative Time



We use the sentiment generated by ChatGPT to generalize the ETS-AAA model in the context of State-Space Models by attributing the role of consumer sentiment to the slope of the local trend (Iezzi & Monte, 2024a; 2024b). We also consider three univariate benchmarks constituted by two ETS models and a MARSS model (Holmes et al., 2024) with univariate observations and two bivariate benchmark MARSS models in which VADER and BERT estimate consumer sentiment.

Table 1 summarizes some information statistics and accuracy measures for Toyota Camry's monthly sales volume forecasting in the US, both with and without the observed proxy signal of consumer sentiment. Regarding information statistics, we achieved the best results using the sentiment proxy derived by ChatGPT. However, regarding accuracy measures, BERT resulted in a strong competitor.

Table 1 - Information parameters and accuracy measures for Toyota Camry US monthly sales volumes: ETS-ANA model, ETS-AAA model, MARSS univariate, and MARSS bivariate with VADER, BERT, and ChatGPT proxy of consumer sentiment

Models	logLik	AIC	BIC	AIC_C	RMSE	MAE	MAPE	SMAPE %	MASE	RAMSSE
ETS-ANA (y)	*-1591.434*	*3212.869*	*3257,623*	*3216.561*	*9440,055*	*8610.602*	*39.914*	*15.902*	*1.669*	*1.389*
ETS-ANA (y)	*-1592.126*	*3218.252*	*3268.973*	*3223.033*	*10809.330*	*10241.120*	*46.714*	*18.321*	*1.985*	*1.594*
MARSS (y) No sent.	*-1567.278*	*314.555*	*3184,319*	*3156.185*	*7291.802*	*6087.463*	*28.754*	*11.831*	*1.180*	*1.073*
MARSS (y) VADER	*-1799.736*	*3623.473*	*3659.276*	*3625.818*	*3678.055*	*3006.139*	*14.435*	*6.630*	*0.583*	*0.541*
MARSS (z) VADER	*-1799.736*	*3623.473*	*3659.276*	*3625.818*	*3678.055*	*3006.139*	*14.435*	*6.630*	*0.994*	*1.232*
MARSS (y) BERT	*-1780.742*	*3585.484*	*3621.287*	*3587.830*	*3542.655*	*2912.880*	*12.455*	*6.508*	*0.565*	*0.521*
MARSS (z) BERT	*-1780.742*	*3585.484*	*3621.287*	*3587.830*	*3542.655*	*2912.880*	*12.455*	*24.351*	*0.965*	*0.985*
MARSS (y) ChatGPT	*-1506.265*	*3036.531*	*3072.334*	*3038.876*	*3321.115*	*2816.744*	*12.629*	*6.231*	*0.546*	*0.488*
MARSS (z) ChatGPT	*-1506.265*	*3036.531*	*3072.334*	*3038.876*	*3321.115*	*2816.744*	*12.629*	*12.809*	*0.780*	*1.059*

Legend: Accuracy measures: logLik=log-likelihood, AIC=Akaike information criterion, AICc=Akaike information criterion corrected, BIC=Bayes Information Criterion, RMSE=Root Mean Square Error, MAE=Mean Absolute Error, MAPE=Mean Absolute Percentage Error, SMAPE=Symmetric mean absolute percentage error, MASE=Mean Absolute Scaled Error, RMSSE= Root Mean Square Error.

4. Conclusions

Our analysis suggests significant potential in using an LLM, such as ChatGPT, both in the unsupervised classification of sentiment in text and the applications of the derived sentiment proxy to predict sales of products/services. Without a pre-trained dataset, we should have applied model-based sentiment systems, such as VADER, with inferior performance. To improve the performances, we would need to train the model using data from a different domain, which presents an evident challenge in generalizing the classification model. On the other hand, we obtained good accuracy by applying BERT based on the Sentiment140 training set, a widely used dataset for sentiment analysis, particularly within the Tweet context. However, this dataset is relatively old and may not reflect current linguistic trends or modern expressions; it may be misfocused on the domain of interest. In cases like this, sentiment classification using generative AI like ChatGPT could be a good solution. Indeed, sentiment classification using an LLM is constantly updated and context-focused. Nevertheless, using ChatGPT requires specific computer programming skills. We think that using ChatGPT to train BERT would be interesting. We are working on this for a forthcoming paper.

Reference

Ali, N., Onur, V. (2024). TurkishBERTweet: Fast and reliable large language model for social media analysis. *Expert Systems with Applications*. 255. 124737. 10.1016/j.eswa.2024.124737.

Assiri, A., Gumaei, A., Mehmood, F., Abbas, T., Ullah, S. (2024). DeBERTa-GRU: Sentiment analysis for large language model, *Computers, Materials and Continua*, 79(3), pp. 4219-4236, ISSN 1546-2218, https://doi.org/10.32604/cmc.2024.050781.

Bogireddy S. R., Dasari. (2024). Comparative analysis of ChatGPT-4 and LLaMA: Performance evaluation on text summarization, data analysis, and question answering. In *15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725662.

Devlin, J., Chang, M.-W., Lee, K. Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding (*arxiv:*1810.04805)

Eysenbach, G. (2023). The role of ChatGPT, Generative language models, and artificial intelligence in *Medical Education: A Conversation with ChatGPT*. *Journal of Medical Internet Education* 9: e46885.

Farhat, F., Silva, E. S., Hassani, H., Madsen, D. Ø., Sohail, S. S., Himeur, Y., Alam, M. A., Zafar, A. (2023). Analyzing the scholarly footprint of ChatGPT: Mapping the progress and identifying future trends. Preprints 2023, 2023062100. https://doi.org/10.20944/preprints202306.2100.v1

Giordano, V., Spada, I., Chiarello, F., Fantoni, G. (2024). The impact of ChatGPT on human skills: A quantitative study on twitter data, *Technological Forecasting and Social Change, 203,*

Harvey, A.C. (1990). *The Econometric Analysis of Time Series, 2nd Edition*, MIT Press Books, The MIT Press, Cambridge, Massachusetts.

Holmes EE, Ward EJ, Scheuerell MD, Wills K (2024). MARSS: Multivariate autoregressive state-space modeling. *R package version 3*. https://CRAN.R-project.org/package=MARSS.

Hou, G., & Lian, Q. (2024). Benchmarking of commercial large language models: ChatGPT, Mistral, and Llama. *Research Square.* (https:orcid.org/0009-0004-6873-5430)

Hutto, C. J., Gilbert, E.E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media* (ICWSM-14).

Hyndman, R.J., Akram, M., Archibald, B.C. (2008). The admissible parameter space for exponential smoothing models. *Annals of the Institute of Statistical Mathematics* 60(2), pp. 407-426.

Iezzi D. F., Monte R. (2024a). Social media effects on sales: Consumer sentiment in a state-space model. In: (eds): M. Misuraca and G. Giordano, *New Frontiers in Textual Data Analysis* (Book Series: *Studies in Classification, Data Analysis, and Knowledge* Organization), *Springer Nature*., ISBN: 978-3-031-55916-7, ISSN: pp. 1431-8814.

Iezzi, D.F., Monte R. (2024b). Building deep-learning consumers' sentiment signals for sales forecasting: A comparative study. In: A.D. D. Longrée,(ed.) *Mots Comptés, Textes Déchiffrés.* p. 419-428, Louvain: Presses universitaires de Louvain, ISBN: 978-2-39061-471-5.

Koonchanok, R., Pan, Y., Jang, H. (2023). Public attitudes toward chatgpt on twitter: Sentiments, topics, and occupations. *Social Network Analysis and Mining*, 14, 1-13. 123389, ISSN 0040-1625.

Kwon, O.H., Vu, K., Bhargava, N., Radaideh, M.I., Cooper J., Joynt, V., Radaideh, M.I. (2024). Sentiment analysis of the United States public support of nuclear power on social media using large language models, *Renewable and Sustainable Energy Reviews*, 200, 114570, ISSN 1364-0321,

Mariani, M.M., Machado, I., Magrelli, V., Dwivedi, Y.K. (2023). Artificial intelligence in innovation research: A systematic review, conceptual framework, and future research directions, *Technovation*, 122, 102623.

OpenAI, T. B. (2023). Chatgpt: Optimizing language models for dialogue. OpenAI.

Varga, M., Albuquerque, P. (2024). The impact of negative reviews on online search and purchase decisions. *Journal of Marketing Research*, *61*(5), pp. 803-820. https://doi.org/10.1177/00222437231190874

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

doi.org/10.26398/asaproc.00117

Modeling virality in social media

Domenica Fioredistella Iezzi[a], Roberto Monte[b], Daniele Pasquini[a]
[a] Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome, Italy
[b] Department of Civil Engineering and Computer Science Engineering, Tor Vergata University, Rome, Italy

1. Introduction

Web virality refers to the rapid acquisition of views through content sharing. The more people share the content on social networks, the higher the chances of it going viral. According to Rogers et al. (2014), information diffusion refers to the process of spreading information over time via communication channels from individuals or organizations to others within a social system. Sharing content within one's network has become a core activity on social media, making these platforms ideal for disseminating information. The importance of social media in information diffusion lies in their ability to broadcast content to extensive audiences. Social networks facilitate the transmission of information through broad connections, allowing it to reach individuals who might not encounter it otherwise. Highly diffused information can indicate its relevance at a given moment, effectively focusing public attention on pressing issues (Nahon and Hemsley, 2013).

In recent years, there has been a growing body of research on viral posts, given their ability to influence social, economic, and political outcomes in contemporary society (Borges-Tiago et al., 2019). Understanding the mechanisms that regulate virality can lead to determining the popularity and engagement of central trends and topics in society (Lopez et al., 2022).

Extensive literature has analyzed the online diffusion of information and electronic word-of-mouth, highlighting how these processes are analogous to viral contagion mechanisms in biology. Social bonding, emotions, and information exchange are factors that may influence this diffusion (Berger, 2014; Babić-Rosario et al., 2020; Raponi et al., 2022). Ismagilova et al. (2021) reviewed the literature, focusing on the reasons that drive users to share news, considering perceptual, consumption-related, personal, and social factors.

This paper aims to identify critical factors that may lead to a highly retweeted message. To achieve this objective, we will examine several variables: 1. the content of the tweet; 2. the sentiment expressed in the tweet; 3. the presence of multimedia elements (such as audio, video, photos, and emojis; 4. the number of followers of the user who posted the tweet; 5. the construction of the hashtag network; 6. the clustering via the Leiden method. Modeling retweeting probability distribution helped predict retweeting dynamics and analyze virality over time.[1]

We have considered a dataset comprising a random sample of 50,000,000 Italian tweets collected from December 1 to December 31, 2020. It is important to remember that in December 2020, Italy was experiencing a very critical health situation due to the COVID-19 pandemic. The government imposed severe restrictions on the country to contain the spread of the virus, dividing Italy into zones (red, orange, and yellow) based on their health risk level, with the most stringent measures applied in the red zones (Di Candia et al., 2024). These restrictions included partial lockdowns, curfews, and the closure of restaurants, bars, theatres, and schools, alongside strict limitations on interregional travel. The so-called ‘Christmas decree’ further intensified these measures during the holiday season, banning travel between regions and limiting social contacts, to curb the potential rise in infection (Nanni and Napolitano, 2024).

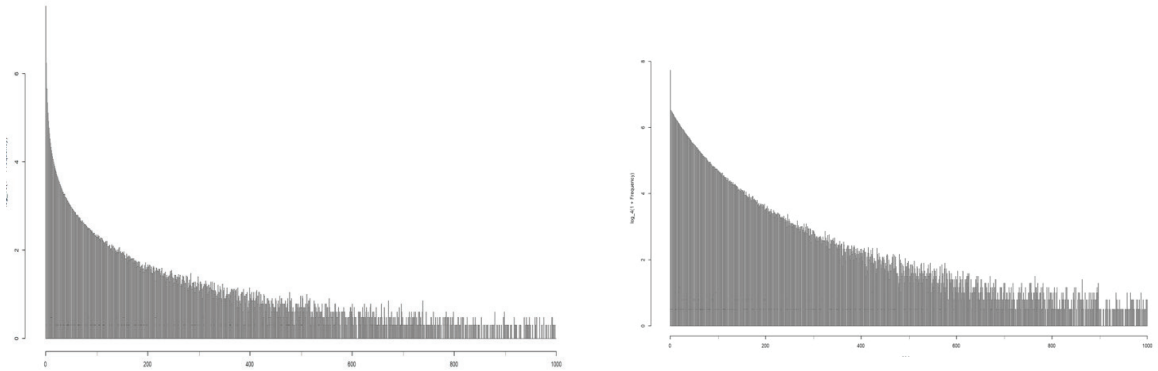
[1] The expression "time of virality" refers to when content becomes viral, spreading rapidly and widely across social networks or the Internet. This concept encompasses the speed at which the content is shared, the moment it reaches its peak diffusion, and the overall duration of its viral cycle.

The remainder of the paper is structured as follows: Section 2 presents retweet distribution, Section 3 presents the main results of our experimentations, and Section 4 illustrates the main conclusions.

2. Retweet distribution

In our sample, 3,136,244 tweets have received at least one retweet. Figure 1 displays two histograms: the one targeted (a) illustrates the distribution of retweeted tweets on a log10 scale, while the other shows a random sample of 365,240 from the zero-inflated beta geometric (ZIBG) distribution, estimated using the method of moments, on a log4 scale. The figures emphasize a strong positive skew and reveals outliers in the right tail, likely representing viral tweets.

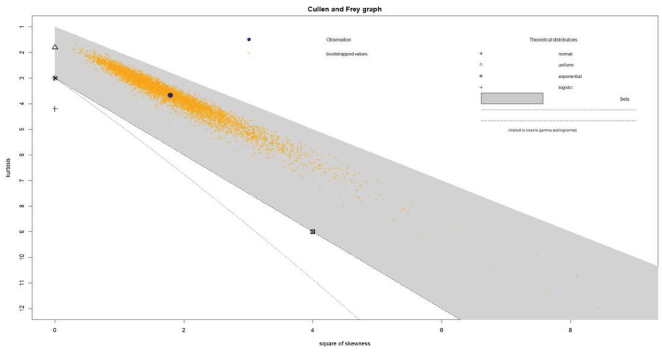
Figure 1. Histogram of 36,523,973 retweeted tweets - $\log_{10}$ scale (a) and histogram of a random sample of size 365,240 from the ZIBG distribution estimated by the method of moments - $\log_4$ scale (b)



(a) (b)

To assess the distribution of the tweets, we have also utilized the Cullen and Frey plot (Cullen and Frey, 1999), a graphical tool to infer the distribution of a data set based on estimating its skewness and kurtosis. Figure 2 presents the Cullen and Frey plot for the sample of 50 million tweets, including those with zero retweets. The plot suggests that the data may follow a mixture of beta distributions.

Figure 2. Full $5x10^7$ tweet sample (zero retweet status tweets included) Cullen-Frey discrete / continuous Plot



To further analyze the distribution of retweets, we propose a straightforward concept: we define the number of failures based on the retweet count of a single tweet. Specifically, a tweet with zero retweets is associated with zero failures, a tweet with one retweet represents one failure, and a tweet

with *n* retweets indicates *n* failures. This concept aligns with the fundamental principle of the geometric distribution, which has a probability mass function (PMF) defined as follows:

$$f(n|p) \overset{\text{def}}{=} p(1-p)^n, \quad n \in \mathbb{N}_0,\ p \in (0,1),$$

where *n* is the number of failures and *p* is the success probability parameter. On the other hand, the tweets with zero failures vastly outnumber all other tweets. Therefore, a step ahead is considered a zero-inflated geometric distribution with PMF given by

$$f(n|\pi,p) \overset{\text{def}}{=} \begin{cases} \pi + (1-\pi)\,p, & \text{if } n = 0, \\ (1-\pi)\,p\,(1-p)^n, & \text{if } n \in \mathbb{N}, \end{cases} \quad p, \pi \in (0,1)$$

which allows us to overweight the probability of zero failures by means of the inflation parameter π. Another issue is tweets are all different from each other in many ways. Hence, we should think of the parameter *p* as depending on the individual tweet. This leads to betting on a mixture of probabilities as the characterizing distribution of the tweets. Formally, we assume that our PMF is given by

$$f(n|\pi,\theta) \overset{\text{def}}{=} \int_0^1 f(n|\pi,p) g(p|\theta)\, dp$$

where *g(p|θ)*, referred to as mixing distribution, is the density of the success parameter *p*. It is worth noting that mixtures of probabilities have been exploited in the context of the so called "buy till you die" models (Fader and Hardie, 2013), which are somewhat similar to ours. In our case, we chose as the mixing distribution the beta density function with shape parameters α, β ∈ R+

$$g(p|\theta) \equiv g(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}.$$

a direct computation shows that our candidate distribution for the tweets is given by

$$f(n|\pi,\alpha,\beta) = \begin{cases} \pi + (1-\pi)\,\frac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)}, & n = 0, \\ (1-\pi)\,\frac{\mathrm{B}(\alpha+1,\beta+n)}{\mathrm{B}(\alpha,\beta)}, & n \in \mathbb{N}, \end{cases} \quad \pi \in (0,1),\ \alpha,\beta \in \mathbb{R}_+.$$

An advantage of the choice of the beta distribution as the mixing distribution is that it is possible to compute the first three moments of the distribution in a closed form. Specifically, we have

$$\mu_1' = \begin{cases} \infty, & \text{if } 0 < \alpha \le 1, \\ \frac{(1-\pi)\beta}{\alpha-1}, & \text{if } \alpha > 1, \end{cases} \qquad \mu_2' = \begin{cases} \infty, & \text{if } 0 < \alpha \le 2, \\ \frac{(1-\pi)\beta(\alpha(\alpha+\beta-1)+\beta(\alpha-2))}{(\alpha-1)^2(\alpha-2)}, & \text{if } \alpha > 2, \end{cases}$$

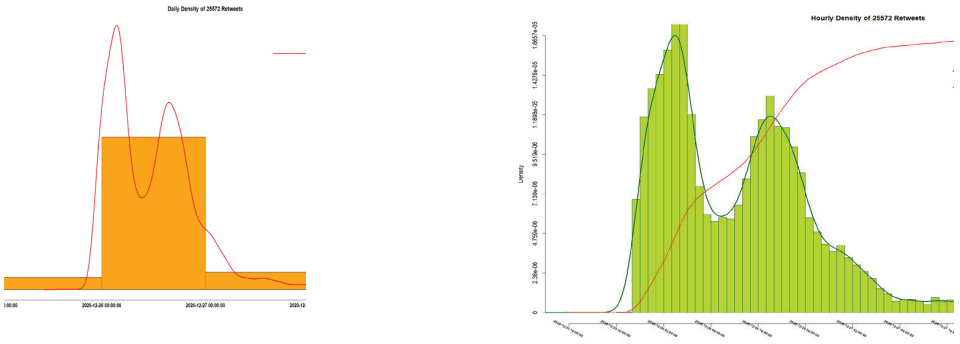
$$\mu_3' = \begin{cases} \infty, & \text{if } 0 < \alpha \le 3, \\ \frac{(1-\pi)\beta(\alpha(\alpha+1)(\alpha+\beta-1)(\alpha+2\beta-1)+(\alpha-3)\beta(3\alpha(\alpha+\beta-1)+\beta(\alpha-2)))}{(\alpha-1)^3(\alpha-2)(\alpha-3)}, & \text{if } \alpha > 3. \end{cases}$$

In addition, we have a closed form for the log-likelihood function, which is given by

$$\begin{aligned} \log \mathcal{L}(\pi,\alpha,\beta|x_1,\ldots,x_n) &= n_0 \log\left(\frac{\alpha+\pi\beta}{\alpha}\right) + (n-n_0)\log(1-\pi) \\ &\quad + \log \mathcal{L}_{\text{beta-geometric}}(\alpha,\beta|x_1,\ldots,x_n) \end{aligned}$$

where $x_1, \ldots, x_n$ are the realized samples and $n_0 \equiv |\{k \in \{1, \ldots, n\} : x_k = 0|$ is the cardinality of the tweets with zero retweets. The closed forms of the moments μ'_1, μ'_2, μ'_3 allow the application of the method of the moments to determine preliminary estimates of the parameters π, α, β, through computational procedures for solving nonlinear-equations, and these preliminary estimates can be used as starting points of the computational methods for maximizing the closed form of the log-likelihood function.

Figure 3. Daily and hourly distribution for a sample viral tweet



Daily density **Hour density**

The time of virality can vary significantly based on the type of content, the platform where it is shared, and the dynamics of the social network involved. For our analysis, we measure virality in hours. For example, one viral tweet from our sample, shared approximately 25,000 times, reached its peak virality within an hour. This illustrates how quickly content can spread on the platform. Using days as a measurement unit is too broad, while using minutes (not reported in this paper) is too narrow. This pattern holds for all viral tweets in our analysis.

3. Discussion

We have analyzed the most retweeted messages, specifically those with at least 1,000 retweets. Our analysis includes a total of 349,491 tweets, which represents approximately 0.69% of all tweets. The messages exhibit strong sentiment connotations: 52% convey negative sentiment, 43% express positive sentiment, and only 5% are neutral. We have conducted a manual sentiment classification because automated dictionary-based, model-based, or machine-learning systems have not yield accurate results. Additionally, 9% of these tweets include at least one emoji.

To examine the content of the tweets, we have employed the latent Dirichlet allocation (LDA) algorithm (Blei et al., 2003). Our analysis has revealed nine distinct topics: 1. Football, 2. Preventive measures for Coronavirus during the Christmas holidays, 3: Behaviors during the lockdown in Italy 4: Encouragement for political and social commitment and love for others, 5. Humor and comedy, 6. Advertising messages, 7. In memory of footballer Paolo Rossi, 8. Politics, 9. Purchases.

We have also constructed a network structure based on the most retweeted messages, which includes the community of users, their followers, and the associated hashtags. To analyze this network, we have evaluated two algorithms commonly used for constructing communities in social graphs: the Louvain and Leiden algorithms (Traag et al., 2019). Both the Louvain and Leiden algorithms aim to maximize modularity measures in community detection. However, we have found that the modularity measure provided by the Louvain algorithm resulted in overly large clusters, which were not particularly useful for our purposes. As a result, we decided to use the Leiden algorithm exclusively.

In building our graph, we have considered two types of interactions, resulting in a heterogeneous network: user-user links for retweets and user-hashtag links when a user connotates a post by a

hashtag. We have then applied the clustering algorithm to the heterogeneous graph, generating mixed communities comprising users and hashtags.

We have utilized various statistical models to investigate the factors influencing retweets (Cameron and Trivedi, 1990; Korosteleva, 2018). Our analysis includes explanatory variables such as sentiment, multimedia elements (e.g., photos, videos, and audio), followers, and network measures (follower, hashtag, and cluster).

Retweet data, which typically consist of count data, often exhibit overdispersion. In such cases, standard Poisson regression models, which assume that the mean and variance of the response variable are equal, may not be suitable. Yet, we have considered the Poisson model as a benchmark alongside two additional models. As shown in Table 1, the regression models for our analysis include Poisson (POIS), negative binomial (NEGB), and zero-truncated negative binomial (ZNEGB). Each model includes eight distinct variations, introducing one regressor at a time. These regressors are categorized into two groups: content-related factors (such as the presence of audio/video, photos, emojis, sentiment, and topics) and the network characteristics of the tweeters (including followers, hashtags, and clusters).

Table 1. Results of regression models: POIS, NEGB, ZNEGB

Akaike Information (AIC)	POIS	NEGB	ZNEGB
audio/video	137769	1677	1814
audio/video + photo	135383	1676	1813
(audio/video + photo + emoje)=multimedia	121399	1653	1793
multimedia+sentiment	112474	1652	1788
multimedia+sentiment+topic	59769	1633	1790
multimedia+sentiment+topic+follower	59014	1634	1635
multimedia+sentiment+topic+follower+hashtag	59014	**1630**	1631
multimedia+sentiment+topic+follower+hashtag+cluster	59016	1632	1633

We have used the Akaike information criterion (AIC) to compare the most performant statistical models and identify those best suited to the data. As shown in Table 1, the negative binomial (NEGB) regression model outperforms all other models when explanatory variables include Multimedia, Sentiment, followers, and hashtags.

Our analyses used R and Python, utilizing custom code and various libraries.

4. Conclusions and future steps

Our analyses, conducted on a large sample of Italian-language tweets collected during the initial phase of the Covid-19 pandemic and the Christmas season of 2020, allowed for a detailed analysis of several characteristics of tweets that were retweeted at least 1,000 times, including multimedia aspects, contents, sentiments, followers, hashtag and clusters. The probability distribution of 3,136,244 tweets that received at least one retweet aligns with a mixed beta density function, showing a peak of virality occurring within one hour.

Different regression models (POIS, NEGB, ZNEGB) indicate that the retweet propagation is influenced by both the content of the message and the structure of the author's network. In sports-related tweets, a distinct community of users emerges who closely follow posts from their favorite athletes. For example, with 110.8 million followers, Cristiano Ronaldo provides substantial visibility for his messages, resulting in a high potential for sharing.

Similarly, influential figures like sports commentator Ibai, who has 14.3 million followers, demonstrate similar dynamics. This highlights how engaging content, combined with a large and active follower base, can significantly increase the spread of tweets.

References

Babić-Rosario, A., de Valck, K, Sotgiu, F. (2020). Conceptualizing the electronic word-of-mouth process: What we know and need to know about eWOM creation, exposure, and evaluation. *Journal of the Academy of Marketing Science,* **48**, pp. 422–448.

Berger, J (2014). Word of mouth and interpersonal communication: A review and directions for future research, *Journal of Consumer Psychology*, **24**(4), pp. 586-607.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**(1), pp. 993–1022.

Borges-Tiago M. T., Tiago, F., Cosme C. (2019). Exploring users' motivations to participate in viral communication on social media, *Journal of Business Rese*arch, **101**, pp. 574-582.

Cameron, A., Trivedi, P. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46(3), pp. 347-364.

Cullen, A., Frey, H. (1999). *Probabilistic Techniques in Exposure Assessment* (1st ed.). New York: Plenum Publishing Co.

Di Candia, F.P., Roberto, N., Iezzi, D.F. (2024). Italian institutional communication in pandemic period: A chronological analysis of prime minister speeches. In: Giordano, G., Misuraca, M. (eds) *New Frontiers in Textual Data Analysis. JADT 2022. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Cham. https://doi.org/10.1007/978-3-031-55917-4_27

Fader, P.S., Hardie, B.G.S. (2013). The Gamma–Gamma model of monetary value. Available: http://brucehardie.com/notes/025.

Ismagilova, E., Rana, N.P., Slade, E.L., Dwivedi, Y.K. (2021). A meta-analysis of the factors affecting eWOM providing behaviour, *European Journal of Marketing*, **55**(4), pp. 1067-1102.

Korosteleva, O. (2018). *Advanced Regression Models with SAS and R* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315169828.

Lopez Y.L., Grimaldi D., Garcia S., Ordoez J., Carrasco-Farre C., Aristizabal A.A. (2022). Artificial intelligence model to predict the virality of press articles. In *Proceedings of the 2022 14th International Conference on Machine Learning and Computing* (ICMLC '22) (pp. 221–228). Association for Computing Machinery, New York, NY, USA.

Nahon, K., Hemsley, J. (2013). *Going Viral*. Polity Press.

Nanni, R., Napolitano, M. (2024). Auditing the 2020 'Christmas decree': Communicating pandemic restrictions during Italy's first 'Covid-Christmas'. *Italian Political Science*, 18(2), pp. 122–138.

Raponi, S., Khalifa, Z., Oligeri, G., Di Pietro, R. (2022). Fake news propagation: A review of epidemic models, datasets, and insights. *ACM Transactions on the Web*, **16**(3), pp. 422-448.

Rogers, E.M., Singhal, A., Quinlan, M.M. (2014). Diffusion of innovations. In E.M. Rogers, A. Singhal, M.M. Quinlan (eds). *An Integrated Approach to Communication Theory and Research*, (pp. 432-448). Routledge.

Traag, V.A., Waltman, L., van Eck, N.J. (2019) From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports,* **9**, pp. 5233 https://doi.org/10.1038/s41598-019-41695-z.

doi.org/10.26398/asaproc.00118

Organizational flexibility measures: Remote working at the Italian National Institute of Statistics

Chiara Limiti, Giuseppina Sola,
Directorate General, Istat, Rome, Italy

1. Introduction

In recent years, terms such as *remote working*, *telework*, and *working from home* have entered common usage and are often used interchangeably. However, it is important to recognize that these terms refer to distinct organizational models, each with its own characteristics and specific regulatory frameworks. Understanding these differences is crucial, as each model offers particular advantages and poses unique challenges. The selection of the most appropriate model should be guided by the need to balance individual well-being and work-life harmony with broader corporate development strategies.

Moreover, the widespread adoption of these flexible work arrangements—spurred by the response of businesses and public administrations to the Covid-19 pandemic—has highlighted the need for statistical measurement of the various forms of remote work and the definition of shared conceptual standards (ILO, 2020).

The objective of this paper is to present an initial analysis of available data on how the Italian National Institute of Statistics (Istat), in accordance with current regulatory provisions, implemented a remote working model during the 2020–2024 period. This model is characterized by a high degree of flexibility and employee accountability, fostering the development of a results-oriented work culture and enabling the reconciliation of personal and family responsibilities with professional demands.

2. An overview of the legislative framework for remote working in the Italian public sector

Although remote working has a long-standing tradition in the Italian legal system—dating back to the so-called third Bassanini law of 1997—remote working in Italy was formally defined by Law No. 81 of 2017, which established the legal basis for its application in public administrations. According to Article 18 of Law No. 81/2017, "remote working is a mode of execution of the employment relationship established by agreement between the parties, including forms of organization by phases, cycles, and objectives, without fixed working hours or a designated workplace, and potentially supported by technological tools to carry out work activities."

Directive No. 3/2017 on remote working, issued by the Minister for Simplification and Public Administration, provided operational guidelines for its implementation and for launching pilot programs.

The Covid-19 health emergency had a significant impact on this framework, effectively ending the so-called "experimentation phase" and transforming remote working into the standard mode of work. The regulatory measures introduced during the pandemic—particularly the renewal of national collective labor agreements (2019–2021), finalized in the post-pandemic period—established remote working as a permanent organizational model, characterized by time flexibility and voluntary participation.

Remote working is a form of work characterized by the absence of time or spatial constraints and organized around phases, cycles, and objectives, established through mutual agreement between employee and employer. It is designed to promote both productivity and a better balance between work and personal life.

Remote work refers to work performed in compliance with attendance obligations derived from standard working hours, typically at a designated workstation outside the office. It can take two main forms: a) Home-based telework, where activities are carried out from the employee's residence; b) Other remote arrangements, such as co-working spaces or decentralized satellite offices.

Remote workers are legally entitled to compensation and employment conditions equivalent to those of employees working on-site, in accordance with collective bargaining agreements. This type of working is governed by a voluntary agreement within the broader employment relationship, allowing employees to choose their working hours and location, within the framework of general arrangements with their employer. Remote work may be defined within the employment contract itself (or through subsequent amendments), and entails predetermined locations, tools, and working hours.

3. Remote working at Istat: Organizational solutions adopted

In accordance with the prevailing regulatory provisions at the time, Istat launched a pilot initiative for remote working in mid-February 2020, following the definition of guiding principles in collaboration with trade unions. Originally intended to last three months, the experiment was preceded by an organizational analysis aimed at identifying tasks within central directorates and services that could be performed remotely—specifically those not requiring continuous physical presence and compatible with data confidentiality and security standards.

The initiative allowed broad participation across the institute, including part-time employees and managers, while excluding shift workers, teleworkers, and probationary staff. The standard limit was four remote working days per month, extendable to six under exceptional or unforeseen circumstances, at the discretion of the director. Notably, the experiment was cost-neutral: Istat was not obligated to supply necessary equipment or reimburse personal IT and internet usage. Although over 900 agreements were signed, only 210 employees actually began working from remote in late February 2020.

The outbreak of the Covid-19 pandemic prompted a shift to "emergency remote working." In early March 2020, Istat quickly expanded participation from 210 to approximately 2,000 employees (Limiti and Sola, 2024). To sustain operations, the institute distributed laptops and connectivity tools, upgraded IT infrastructure, deployed relevant software, and provided technical support through multiple online channels. Within 10 days, about 1,700 virtual desktop infrastructures (VDIs) sessions and 150 virtual private networks (VPNs) were configured, alongside the setup of additional servers. This enabled Istat to maintain statistical production and dissemination despite halting data collection in early-affected regions from February 26, 2020.

As the pandemic worsened, Istat closed all territorial and decentralized Roman offices, restricting access to the central headquarters to a limited number of essential personnel (around 30 units). Employees in telework arrangements were exempted from returning on-site, all travel missions were suspended, and all meetings transitioned to online platforms.

Following the "relaunch decree" (Decree-Law No. 34 of May 19, 2020, Article 263), which called for public services to adapt to the progressive reopening of economic activities, Istat upheld protective measures for vulnerable staff, even beyond legal mandates. It also planned a cautious return to in-person work by reopening peripheral offices with occupancy limits and monthly rotation schedules. To minimize travel during peak hours, entry and exit times were made highly flexible, with a one-minute presence window set between 12:44 and 12:45 for levels IV-VIII employees. A three-hour daily availability window was also introduced to ensure connectivity and protect the right to disconnect.

From October 2021 through 2022, following a decree by the President of the Council of Ministers (September 23, 2021), in-person work was reinstated as the standard mode in public administration. Istat adopted a prevalence model, requiring that 51% of working days be performed

on-site, with up to 20 remote working days allowed per two-month period. Remote working could be scheduled as 16 full days and up to 4 half-days. Later adjustments recognized "fragile" and special teleworkers as eligible for 100% remote working.

In 2023, the flexible model was consolidated with the introduction of decentralized work, enabling staff to operate from any Istat office across Italy. This was supported by digitization efforts and the institute's nationwide presence, with at least one office in each region. That year, the remote working cap was revised to 10 days per month, including up to 2 half-days.

In 2024, having moved beyond the emergency phase, Istat implemented a directive of the Minister for Public Administration (December 29, 2023) and began aligning with Title III of the national collective agreement for the education and research sector (2019–2021). This led to the introduction of an "enhanced" remote working model. While ordinary remote workers are allowed up to 20 days per two-month period, enhanced remote workers—those with serious and otherwise irreconcilable health or family needs, including long commutes—are permitted up to 24 days.

4. Analysis methodology and data source

To analyze the uptake of remote working, administrative attendance records were examined. Specifically, justification codes from the URBI personnel management system were analyzed for the 2020–2024 period (Limiti, 2024). A unified dataset was constructed based on monthly attendance logs and included all new codes introduced for remote working. Teleworkers, special teleworkers, and shift workers were excluded due to their distinct arrangements.

To enable comparison with other managerial data such as hours worked, the following decisions were made:

- Remote working justification codes were counted in the month they were used, regardless of whether internal regulations defined use on a monthly or bi-monthly basis. This may obscure the effects of seasonal absences (e.g., during holidays).
- Mixed-mode remote working days were recalculated using the contractual daily working time (7 hours and 12 minutes), enabling accurate computation of hours worked (Sola, 2025).

The distribution of remote working was analyzed by gender and fragile worker status. Future work will integrate additional administrative data, such as age, length of service, qualification, and responsibility level, to deepen the analysis.

5. Preliminary findings

The majority of Istat employees have been engaged in remote working practices. During the analyzed period (2020–2024), the quota of staff who performed at least one day of remote working in December consistently exceeded 78% of the total workforce, with a marked prevalence among female employees.

For instance, in December 2024, 92% of women (1,011 individuals) and 87% of men (644 individuals) had worked at least one day in remote working mode. These figures represent an increase of 4% and 3%, respectively, compared to the same month in 2023. Participation rates in December 2024 closely resembled those recorded during the peak of the COVID-19 pandemic, when remote work was mandated by law. In December 2020, 96% of female employees (1,177 individuals) and 88% of male employees (774 individuals) had engaged in at least one day of remote working.

Throughout the entire 2020–2024 period, the number of women working remotely never dropped below 889 (corresponding to at least 80% of female staff in service), while the number of men never fell below 539 (corresponding to at least 75% of male staff).

In terms of total days, following a peak in 2020–2021, remote working days declined to just over 188,000, with women accounting for 62% of those days.

Annual averages suggest employees used approximately 170 remote working days per year,

declining to about 110 during the normalization period. Gender differences in usage were minimal in 2024: men averaged 115 days, women 117, suggesting that remote working has been used equitably as a work-life balance tool.

Figure 1 – Smart working days used by employees by gender and year (percentage values)

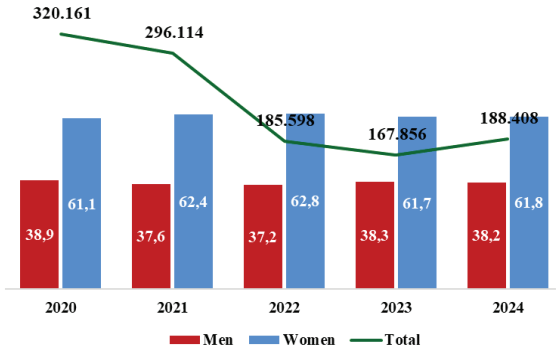
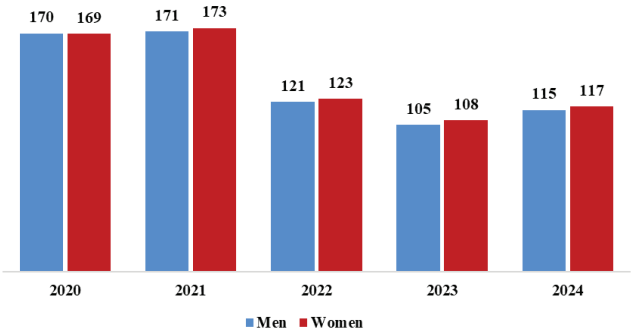


Figure 2 - Smart working days used by employees by gender and year (average)



When analyzing average remote working days by gender, no significant differences emerge: in 2024, men worked an average of 115 days remotely, while women worked 117 days on average (Figure 2). This suggests that remote working at Istat was experienced as a tool for work-life balance for all employees, and not primarily as a benefit for women, who are traditionally more involved in caregiving responsibilities.

In 2022, following the introduction of the "fragile workers"[1] category, Istat exempted these employees from returning to the office, granting them 100% remote working without monthly or bimonthly limits. Fragile employees who worked at least one day remotely accounted for nearly 9% of total staff in 2022 and 2023, a figure that declined to about 3% in 2024. The total number of remote working days for this group reached approximately 11,000 in 2024. Over the entire period considered, more than 64% of these days were worked by women (Figure 3), although on an annual average, fragile male employees tended to work more remote working days than their female counterparts (Figure 4).

Figure 3– Smart working days used by fragili employees by gender and year (percentage values)

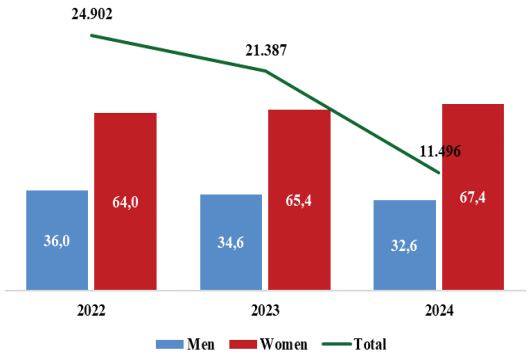
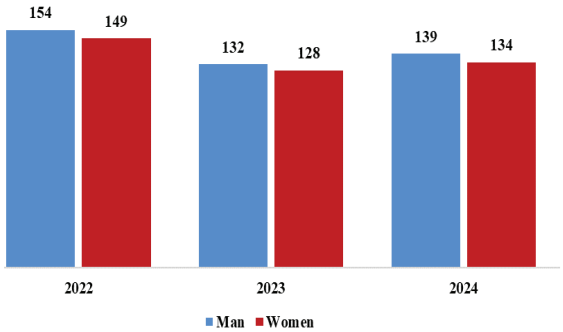


Figure 4– Smart working days used by fragili employees by gender and year (average)



[1] According to the definition stipulated in the decree issued by the Ministry of Health on 4 February 2022, the term 'fragile workers' encompasses individuals exhibiting a pronounced immune system impairment, patients afflicted with a minimum of three or more concurrent serious ailments as delineated by the aforementioned ministerial decree, and those with a documented exemption from vaccination.

6. Potential critical issues

The remote working model adopted by Istat represents a major shift from the traditional work paradigm, placing emphasis on the integration of four key dimensions: people, processes, connectivity, and technology. Its most distinctive feature is a high degree of flexibility. However, for this model to evolve into an effective and sustainable organizational framework, it must also address the critical issues that such flexibility can introduce.

First, location flexibility—the absence of an obligation to declare one's working location—calls for strategies aimed at fostering and maintaining social relationships, which are vital for the development of human capital. While working remotely reduces commuting time and costs and allows employees to personalize their work environment—leading to greater comfort, concentration, and satisfaction—it also risks fostering isolation. To counteract this, organizations must identify solutions or create new opportunities for socialization that preserve the sense of shared purpose, teamwork, and interpersonal exchange.

Second, temporal flexibility—with no fixed working hours except for designated windows of availability—requires both regulations and technological tools to safeguard the right to disconnect. While this flexibility acknowledges individual differences in productivity cycles and allows for a better balance between work and personal life, it also blurs the boundaries between professional and private time. Without clear rules, there is a risk of extended work commitments beyond office hours, on weekends, holidays, or even during sick leave. Ensuring proper boundaries is therefore essential to protect employee well-being.

Third, organizational flexibility—granting autonomy in managing one's own tasks once objectives are defined—demands clear guidance and support from managers to prevent performance inconsistencies and professional isolation. This autonomy encourages accountability and a results-oriented culture, fostering motivation and satisfaction. However, it requires that managers possess strong planning skills and the ability to set concrete and realistic goals. Moreover, recognizing and valuing the specific skills of each employee is essential to ensure that tasks are appropriately matched to individual strengths. When properly implemented, this approach can yield significant benefits, including higher productivity, stronger engagement, and enhanced job satisfaction.

References

International Labour Office. (2020). *Defining and Measuring Remote Work, Telework, Work at Home and Home-Based Work*. Geneva: ILO.

(https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@stat/documents/publication/wcms_747075.pdf)

Limiti, C., Sola, G. (2024). Il lavoro agile nell'Istat – Istituto nazionale di statistica. Poster presented at the 15th National Conference on Statistics, Rome.

Limiti, C. (2024). *Il lavoro agile – Report di monitoraggio gennaio–settembre 2024*. Technical note, Istat, Rome.

Sola, G. (2025). *Lavoro agile nell'Istat – Istituto Nazionale di Statistica. Anni 2020–2025*. Technical note, Istat, Rome.

doi.org/10.26398/asaproc.00119

Improving information on labour demand for decision-making by integrating traditional surveys, administrative sources, and web-based data

Annalisa Lucarelli, Alessandra Righi, Roberta Rizzi
Italian National Institute of Statistics.

1. Introduction

The information gap that has characterised the demand side of the labour market compared to the supply side is gradually closing thanks to the use of new sources of information, administrative and web data, in addition to traditional sources.

This paper outlines possible scenarios for integrating traditional and innovative sources to deepen the understanding of labour demand for decision-making. We analyse both the characteristics and the limitations of existing sources, as well as some labour demand measures derived from the different sources. A distinction is made between sources covering **met labour demand** (i.e. occupied job positions) and **unmet labour demand** (i.e. vacant job positions). As far as met labour demand is concerned, the Istat OROS survey (based on Social Security Institute administrative data) covers the stock of job vacancies while the Compulsory Communications (CCs, Ministry of Labour administrative source) cover the flows of events of activation, termination, extension or transformation of each employment relationship. For the unmet labour demand, Istat's official quarterly survey estimates job vacancy rates, broken down only by NACE Rev. 2 economic activity sections. The Online Job Advertisements (OJAs) is an innovative web-based data source with various limitations in the completeness of the activity sectors considered and data quality issues but provides detailed information on the search for workers by enterprises, broken down by strategic variables not available in the Istat official survey. We test the ability of this new source to correctly describe the dynamics of the labour demand by comparing the measured trends with those observables using the two sources traditionally used for the purpose. We analyse whether these three sources converge in providing indications on labour demand trends usefull for decision-makers despite the different data quality and definitions. We compared both the quarterly year-on-year changes in the official Istat vacancy rate with the changes estimated from the OJA rates, and the stock of OJAs for each month with the employment inflows derived from the CCs in the following months with different lags. If the indications were similar, the timeliness of the web-based source could be used to provide warnings of business cycle trends and turning points.

2. Data sources on labour demand

In the context of the Italian official statistics covering the demand side of the labour market, there are the Compulsory Communications (CCs) of the Ministry of Labour[1] – which provides quarterly information on flows of activations, terminations, and jobs communicated by the employer – and the Istat job vacancy surveys (VELA/LES) – which supply quarterly information on vacancies –. A vacancy is defined as an unoccupied workplace that the employer wants to have filled, taking active action to hire a worker. In addition to the traditional official sources mentioned above, there is the big (web) data source represented by the Online Job Advertisements (OJAs),

[1] Compulsory Communication are not open data. However, Istat receives these data according to a regulation (Legislative Decree 322/1989, 196/2003) that obliges administrative bodies to provide the data.

collected by the European Centre for the Development of Vocational Training[2] (CEDEFOP-Eurostat). OJAs refer to advertisements published on several web sources such as job portals, company sites, social networks, employment agencies, job search engines, online newspapers and public employment services. CEDEFOP data are extracted by means of the three following ingestion methods: scraping, crawling and APIs usage.

The characteristics of the three different sources of labour demand are briefly summarised in Table 1. Although the three sources provide interrelated variables, they have many other different characteristics. The main difference lies in the nature of the source: administrative the first (CCs), survey the second (VELA), and big data the third (OJAs). Concerning the subjects required to provide data, while in CCs and OJAs are all employers belonging to sections A-U of the NACE Rev. 2, in the survey VELA enterprises with at least one employee in the private non-agricultural sector (sections B-S) provide data. CCs and VELA have the same coverage in terms of employment type (regular and para-subordinate employment, e.g. individuals who are legally self-employed but who are often "economically dependent" on a single employer so as to be assimilated to an employee) while they do not include self-employment, which is instead covered by OJAs. Another major difference is the unit of analysis: respectively employees and para-subordinate jobs (Linked employer-employee database-LEED) (CCs), enterprises (VELA), and job advertisements (OJAs). The level of detail of the variables collected by the three sources also differs significantly. OJA data offer detailed information regarding the characteristics of the job (e.g. occupation, location, type of contract, working time, and salary); the characteristics of the employer (e.g. economic activity sector); the job requirements (e.g. education, skill, and experience); and information on the advertisement (e.g., job portal and publishing and the expiring date of the ad). There is no exact correspondence between an advertisement and a job because each advertisement can relate to different job positions or cannot report the number of positions requested. Vacancy data from VELA/LES are only provided broken down by NACE Rev. 2 economic activity sections and at the national level, while CCs are available by occupation, NACE Rev. 2 economic activity sections, duration of employment (for fixed-term jobs), and many other variables referring to the worker and the employer.

CCs and OJAs collect daily data disseminated at the regional and national level, while the collection in VELA is quarterly and at the national level. The best timeliness is given by OJAs (30 days after the end of the reference period), CCs disseminate data 50 days after the reference quarter, while VELA provides a flash estimate and a final estimate, respectively, 45 and 70 days after the reference quarter.

Table1: Characteristics of traditional and innovative sources on labour demand

		Compulsory Communications (SISCO system in Ministry of Labour)	**Survey on job vacancies an worked (VELA/LES)**
Data owner		Ministry of Labour	Italian National Institute of S
Type of source		**Administrative** source (statistically treated) based on the telematic transmission of communications relating to events of activation, termination, extension and transformation of employee and para-subordinate employment relationships by employers (form UNILAV).	**Survey** source, based on a sa around 29.000 enterprises. D collected separately for enter with 1-499 and those with at 500 employees, respectively Istat VELA survey and the ce monthly Istat survey on large enterprises.

[2] Results based on online job advertisements (OJAs) are available in the Skills-OVATE dashboard at the link https://www.cedefop.europa.eu/en/tools/skills-online-vacancies. This tool is powered by Cedefop's and Eurostat's joint work in the context of the Web Intelligence Hub. The access to CEDEFOP OJA data is organised by Eurostat (Web Intelligence Hub team: ESTAT-WIH@ec.europa.eu).

Table 1 (cont'd): Characteristics of traditional and innovative sources on labour demand

Subjects required to provide data	Employers: both companies and public institutions (excluding the Armed Forces) and households (for domestic work), residents in Italy. In particular: Private employers (excluding temporary employment agencies in form UNILAV), domestic employers, and Public Administration. NACE Rev. 2 sections A-U.	Employers: enterprises with at least one employee in the private non-agricultural sector, resident on Italian territory. NACE Rev. 2 sections B - S. Public institutions in public administration, education, and human health (sections O, P, and Q) are not fully covered.	Employers: enterprises, public institutions (except Armed Forced), and households (for domestic work). NACE Rev. 2 sections A - U.
Coverage	Regular and para-subordinate employment are included, while self-employment is excluded. Temporary work and, partially, the maritime sector are excluded. Top management figures of public and private companies) are excluded.	Jobs for employees (for whom social security contributions are paid) and for apprentices. Managers are included, except for the top management figures. Staff with project and collaboration contracts and self-employees are excluded.	Jobs for employees and apprentices and self-employees. Managers are included
Statistical unit of analysis	Employees and para-subordinate jobs of Italian/ foreign citizens. Events of activation, termination, extension and transformation of employee and para-subordinate employment relationships.	Enterprises	Job advertisements
Statistical variables and level of detail	• Quarterly flows of activations and terminations of employment relationships and quarterly flows of jobs. • Average number of activation or termination events per worker. • Workers affected by at least one activation or termination event. Data are broken down by occupation (permanent/fixed-term), NACE Rev. 2 economic activity section level, duration of employment (for fixed-term jobs).	• Stocks of job vacancies and occupied posts. • Inflow and outflows of occupied posts. • Hours worked. Data are broken down by NACE Rev. 2 economic activity section level	• Job advertisements published on the World Wide Web. Data are broken down by occupation (ISCO-08), skill (ESCO), territorial area (LAU and NUTS 2021), economic activity sector (NACE Rev.2). • Education required, contract type, working time, experiences required (in years), gross salary range offered.
Frequency of data collection and territorial detail	Data collected continuously (at each activation, transformation, extension, and termination event) Disseminated quarterly indicators at the national and regional level.	Quarterly data at the national level	Daily data disseminated at the (municipalities) region level.
Timeliness	50 days after the reference quarter.	Estimates 45 (flash estimate) and 70 (final estimate) days after the reference quarter.	Micro-data are released one month after the end of the reference period.

3. The comparison of the labour demand dynamic among sources

We compared the dynamics of labour demand described by the OJA trends with the trend registered by the other two official sources, the VELA/LES survey and the administrative CCs.

VELA/LES measures quarterly job vacancies (opened on the last day of each quarter). The

indicator disseminated is the official job vacancy rate, which measures the number of vacancies for which search/selection activities are in progress out of 100 vacancies that are either already occupied or for which a recruitment process is taking place. As described in the previous section (, the VELA/LES indicators are only available at the national level and by Nace Rev. 2 economic activity sections. CCs provide the level of flows into and out of employment, broken down by occupation and available on a monthly basis. While for comparison between the VELA/LES official vacancy rate an OJA rate has been constructed using the same formula as the official vacancy rate, the consistency between OJAs and CCs has been assessed by testing different time lags of CCs to take into account the unknown time it takes ads to turn into new employment. Thus, the stock of OJAs at the end of each month is compared to the monthly activation flows of the CCs with a lag of one to four months. As the official vacancy rate is only available by economic activity, we broken down the OJA rate by economic activity; the monthly OJA stock is disaggregated by occupation, as this information is provided by the CCs.

The year-on-year changes of all four measures (quarterly JV and OJA rate, monthly OJA stocks and CC flows) have been taken into account to analyse the dynamics of labour demand, calculated by comparing the measure of each quarter (or month) with the same quarter (or month) of the previous year. The period considered for the analysis is quite short, from the first quarter of 2022 to the third quarter of 2023, due to quality problems in the OJA source in terms of the temporal stability of job portals/websites from which the OJA data are extracted.

Web-based data needs an accurate evaluation of the data quality before use, differently from other sources (Beręsewicz and Pater, 2021 and Branka et al., 2022). The main quality issues related to OJAs include both representativeness and coverage. OJAs over-represent certain job positions (e.g. for workers with higher levels of education and in large companies), while under-representing others (e.g. seasonal workers in small enterprises in Italy). In terms of coverage, job portals may not be fully covered by the data ingestion activities and, even after collecting all the advertisements published on all existing online portals, the vacancies not advertised online are not covered. Over-coverage could be caused by delays in communication between companies and portals and the consequent presence of expired vacancies on portals. In addition, OJA may not represent a vacancy in the official survey as a company may only be trying to test the market. However, it is mainly the presence of duplicates that leads to problems of over-coverage. Deduplication is a fundamental step in the processing of OJA data to obtain a high-quality set of results. The CEDEFOP team divided this task into three distinct parts: 1) physical deduplication, during the data ingestion phase multiple advertisements from the same URL are considered duplicates: 2) logical deduplication, during the pre-processing phase, the deduplication is based on text similarity; 3) comparison of structured fields resulting from the information extraction phase, at the end of the processing phase the resulting ads are compared by various characteristics. If they share the same values for these fields with another ad, different ads from the previous two phases are also considered duplicates and merged into a single ad.

3.1 Vacancy rate from VELA/LES official survey and OJA

Similarly to the vacancy rate, which is constructed by relating vacancies to the sum of vacancies and occupied posts, the OJA rate was constructed as the ratio of OJAs to the sum of OJAs and occupied posts. The OJAs have been considered in terms of stocks at the end of the quarter, while the number of occupied posts of the quarter in the denominator were taken from the other Istat survey, the OROS survey.

Figures 1 shows the year-on-year changes of the OJA rate and JV rate by the two main NACE Rev. 2 economic aggregate sectors, industry and service respectively, and by two specific economic activity sections, construction and information and communication respectively.

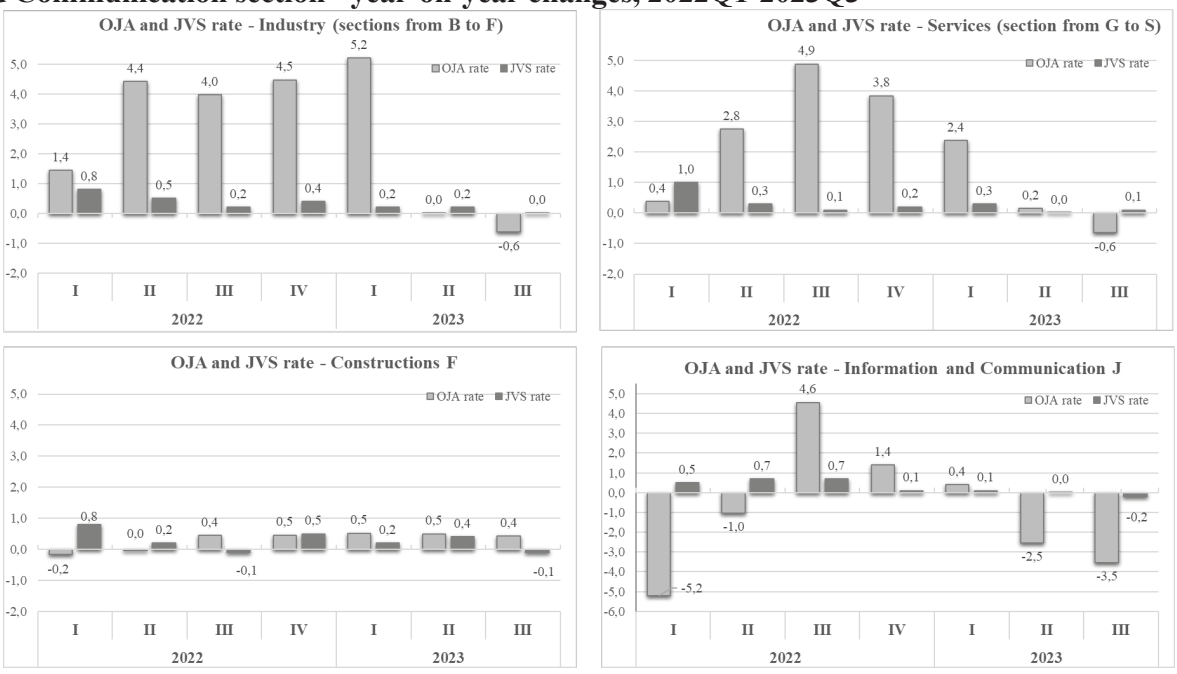
We looked at these two specific economic sections in order to check the different robustness of the OJA data in activities with very different characteristics, also in terms of the use of online advertising by enterprises as a channel for recruiting new staff. The construction sector should be

characterised by a lower use of the online channel in favour of more traditional search channels (e.g. word of mouth). This is due both to the type of enterprises that mainly characterise this sector (presence of many very small enterprises) and to the type of workers sought. In the information and communication sector, enterprises are expected to make greater use of online advertising, due to both the nature of the activities in this section and the type of personnel sought.

The evidence shows that, in both aggregate sectors, the trend changes have the same sign for the rates throughout the period of analysis, except for the third quarter of 2023, when they show a different trend, negative for OJA and stable or slightly positive for JV. Even if the two indicators provide the same positive economic signal in the central period of analysis, the levels of change are different and much more pronounced in the case of OJA, except in the last two quarters. This can be attributed to several factors, the main ones being the difference in definition between vacancies and ads, the limitations in the representativeness of the OJAs, and the improvement in the processing of ads, particularly in the deduplication and classification phase. The similarity of the levels of change over the last two quarters may indeed be due to continuous improvements in the OJA data quality made by the CEDEFOP team.

The evidence at section level does not appear to be as expected. In the construction section, a higher discordance between OJA and JV rates was expected due to the greater use of traditional search channels in this section. On the other hand, a higher concordance was expected in the information and communication section, due to the assumed higher use of the online channel. Misclassifications of ads at a more detailed level by economic activity section may have played an important role in these results. The effects of misclassification errors are less relevant when looking at more aggregated levels of activity.

Figure 1: OJA vs JV rate in Industry and Service sector and in Construction and Information and Communication section - year-on-year changes, 2022Q1-2023Q3



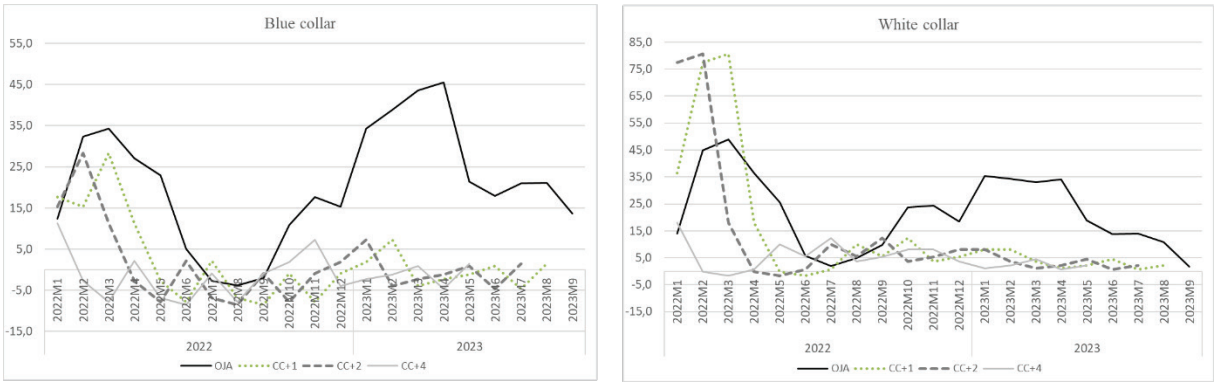
Source: Istat VELA/LES and CEDEFOP OJAs.

3.2 OJA stocks and CC activation flows

We relate now the dynamics of job advertisements (OJAs source) to that of labour activations (CCs source). After having examined the totality of contracts, to better consider the different characteristics of the two sources, our attention is turned to fixed-term contracts. We examined also the occupational qualification, according to the national classification CP2011 and from 2023,

CP2021, the result of a revision of the previous version (CP2011) and further alignment with the International Standard Classification of Occupations (ISCO08). The variable, at group level (first digit), was re-aggregated in the following 4 groups: 1. High-skilled white collar; 2. Low-skilled white collar; 3. High-skilled blue collar; 4. Low-skilled blue collar. The most interesting results emerged for total white collar. The year-on-year changes from January 2022 to September 2023 in the stocks of OJAs on the last day of the month were compared with CCs lagged by one, two, and four months not to make the representation in the graph difficult to understand (the month for CCs is that of the activation date), separately for total blue and white collars (see Figure 2). The dynamics of the monthly OJAs are in line with those of the CCs, taking into account a lag of one to two months, especially for white collars. OJAs seem to show a quite different dynamic to the CCs four months later, in fact they should lead to new employment before the fourth month, as in that month they expire from the portals regardless of whether the position has been filled. The higher level of consistency observed for white-collar workers in the two sources could be due to the fact that OJAs better represent job positions for workers with a higher level of education and qualification (e.g. managers, professionals, technicians, and associated professionals included in white-collar group).

Figure 2: OJA monthly stock vs CC activations flows Blue and White collar, year-on-year changes, January 2022 - September 2023



Source: Ministry of Labour CCs and CEDEFOP OJAs.

4. Open issues and future work

As our goal is to cover better whether the positive indications of convergence between the rates and measures taken from OJAs compared to those of the sources traditionally used are truly promising, we will extend our comparisons considering other key variables of the labour market (business sector, business size, etc.). Unfortunately, the short length of the period makes not possible the use of cointegration analysis to better understand the relationship between the series of different sources.

References

Beręsewicz, M., Pater, R. (2021). Inferring job vacancies from online job advertisements, *Statistical Working Papers*, Eurostat.

Branka, J., Kvetan, V., Napierala, J. (2022). From the online job advertisements to official statistics – The aspects of quality assurance, in *Book of Abstracts of the European Conference on Quality in Official Statistics - Q2022*, Vilnius.

doi.org/10.26398/asaproc.00120

Early detection of university dropouts: An experiment using video-lectures access data

Maria Cristiana Martini
Department of Communication and Economics, University of Modena and Reggio Emilia, Reggio Emilia, Italy.

1. Introduction

University attrition is one of the most complex and adverse events in students' career. The percentage of students who do not complete a bachelor's degree program within its theoretical duration ranges from less than 40% in Israel and the United Kingdom to 80% or more in Colombia, the French Community of Belgium and Italy (OECD, 2022). If we enlarge the time interval to three years after the theoretical duration, the non-completion rate still ranges from 15% in United Kingdom to 50% in Brazil. In Italy 47% of students who entered a bachelor's programme did not graduate three years after the theoretical duration, and 32% are no more enrolled in tertiary education (OECD, 2022).

Attrition is a potentially disastrous event in a student's life, and it also impacts negatively the university. Early identification of students who are prone to abandon their studies allows specialized professionals to implement action plans for retention. Machine learning techniques have widely been used to predict university dropouts (see, e.g., Agrusti et al., 2019 for a review). Student dropout predictions are usually made on the basis of features that depend both on the determinants of university dropouts (see Aina et al., 2022 for a review) and the availability and timeliness of each information. Most predictions are then based on administrative data, like personal information, family background and study career (school background and academic performance), that are costless and timelier, while ad-hoc surveys on involvement and intentions can be more precise but are expensive and difficult to obtain. As students advance through their curriculum, more data are available, but dropouts are less frequent, and interventions become less effective.

The availability of online courses can provide a further source of possible features to predict dropouts, since the video-lectures are often accessed through individual login credentials and these operations produce large log datasets, giving information on time and frequency of video-lectures fruition.

The aim of this paper is to integrate data on video-lectures usage during the first year at university in the prediction of students' dropout, since these data are automatically collected, easily available, and offer an insight on study methods and dedication. Section 2 describes the data and the methodology; Section 3 presents some results, while conclusions are drawn in Section 4.

2. Data and methods

Classes at the Department of Communication and Economics of the University of Modena and Reggio Emilia have been recorded and made available to students starting from September 2017, when the ONELab project was launched (Furini et al., 2020); face-to-face classes were regularly held (with the exception of the pandemic period), but most students accessed the video-lectures to supplement their preparation and review some difficult passage.

The log data is a large dataset where each record corresponds to an access to the ONELab platform, and reports information on the day and time of the access and the accessed course; these data have been aggregated to gather information on the time, type and frequency of accesses for each student.

The data for this study integrates two data sources: log data on video-lectures fruition in the first

months of the first year (from September to the end of May) and administrative data on students' demographics and study career. Data include all the freshmen of the five degree courses of the Department in the academic years 2017/18 and 2018/19, except those who withdrew before June in the enrolment academic year. This exclusion has a double reason: first, the importance of data on video-lectures fruition and academic achievement on dropouts would be magnified, since early dropouts do not pass exams and do not watch video-lectures; furthermore, early dropouts are not interesting to predict, since they are officially known before the data to forecast them is available.

After the deletion of early dropouts, the dataset consisted of 2,269 observations (1,058 enrolled in 2017, 1,222 in 2018), but five were discarded: one died few months after the enrolment, and two couples of homonymous students were enrolled to the same study program, making it difficult to correctly merge their administrative data to the log data. The reason for choosing students enrolled in 2017 and in 2018 is twofold: avoiding the pandemic period, when lectures fruition was completely upset, and including students who have had plenty of time to graduate, so that dropouts' definition becomes simpler.

The first encountered problem, in fact, is the difficulty to define dropouts in an unambiguous and consistent way. Administrative data recorded 395 official withdraws occurring after May in the first University year out of 2264 observations (17.4%), but these were only a part of all the actual dropouts, since many students simply stopped attending classes and paying taxes, but do not withdraw at the administrative office. However, students who stop paying taxes do not always express a final will to dropout: some students might pay late, or might interrupt their studies for a period, due to economic or family problems. After analysing the enrolment status in each academic year up to 2022/23, students have been defined as dropouts if they did not graduate, and they did not enrol to university for at least the last two years: these are 649 cases (28.7%), while the non-completion rate (dropouts and still enrolled students) is 44.3%, and both data are consistent with the OECD (2022) data for Italy.

The final dataset included pseudo anonymized data resulting from the merge of the administrative data and the aggregated log data; besides the outcome variable on dropouts, this dataset included a large number of students' characteristics which can be useful features in a machine learning prediction:

- Basic features, i.e. personal information, family background and school background: gender; age; nationality; specific learning disorders; place of residence; tax area; high school type; high school grade; selected degree program. These features are already available when a student enrolls at the university and can be immediately used to predict dropouts.
- Academic performance in the first semester: number of ECTS, number of passed exams and average grade after the first semester. These features are available after the end of February, when the winter exams session is over.
- Video-lectures usage[1]: total number of accesses; number of courses accessed at least 1, 5, or 10 times; month of first access; number and percentage of accesses per time slot; number of accesses to the favorite course and in the favorite time slot. The availability of these features depends on the time period one wants to consider for video-lectures fruition: this analysis takes into account video-lectures usage between September and May, so these features are available from the month of June. Considering a shorter period implies that data will be available earlier, but will likely be less informative, while a longer observation period implies more informativeness but a delayed data availability.

Some data pre-processing operations have been carried out before data analysis. The dataset has been split in an 80% training data, and a 20% testing data. Although there are very few missing

[1] Since early dropouts are excluded from the dataset and only dropouts occurring after May are considered, the video-lectures usage is limited to the accesses before June.

data, these have been replaced by means of k-nearest neighbours (K-nn) technique. All the features have then been standardized, and One-Hot encoding procedure has been applied to categorical variables. The response variable is dichotomous (dropout/non dropout); hence the prediction regards a binary classification problem. Alternative machine learning algorithms and models have been applied, among the most commonly used in analogous classification problems:

- Decision trees according to the original Classification and Regression Trees (CART) algorithm that uses binary splits (see Breiman et al., 1984);
- Random forest (RF), that is an improved version of CART generated by bootstrapping on the training sample and averaging over the multiple versions of the tree (see e.g. Breiman, 2001);
- K-nearest neighbours (KNN), a basic non-parametric supervised learning method (see Cover and Hart, 1967);
- Neural networks (NN), a machine learning model inspired by the functioning of the human brain, that involves one or more levels of hidden layers; here a simple feed-forward neural network with a single hidden layer is applied (see, e.g., Ripley, 1996);
- Naïve Bayes (NB), a simple linear probabilistic classification algorithm that utilizes the Bayes rule and assumes conditional independence of the features (for a discussion on its optimality under wide violations of the assumption see, e.g, Domingos and Pazzani, 1997).

Each model was estimated on the training dataset, and the hyperparameters were tuned based on a maximization of the area under the ROC curve. Then, each model is assessed on the testing data, with a 10-fold repeated cross-validation. Pre-processing and analyses have been carried out with the *caret* package of R (Kuhn, 2008).

3. Results

Accuracy, sensitivity and specificity are calculated as averages of the measures obtained for each attempt and are reported in Table 1.

Table 1. Accuracy, sensitivity and specificity for the alternative classification methods using different feature sets.

Feature set	Performance measure	CART	RF	KNN	NB	NN
Basic features	*Accuracy*	0.692	0.724	0.677	0.659	0.728
	Sensitivity	0.374	0.503	0.303	0.006	0.503
	Specificity	0.859	0.838	0.872	1.000	0.845
Basic features and academic achievements	*Accuracy*	0.768	0.774	0.721	0.675	0.792
	Sensitivity	0.523	0.574	0.419	0.052	0.600
	Specificity	0.896	0.878	0.879	1.000	0.892
All features	*Accuracy*	0.770	0.790	0.730	0.704	0.777
	Sensitivity	0.548	0.561	0.432	0.600	0.580
	Specificity	0.886	0.909	0.886	0.758	0.879
ONELab usage only	*Accuracy*	0.693	0.695	0.675	0.664	0.684
	Sensitivity	0.445	0.361	0.394	0.465	0.400
	Specificity	0.822	0.869	0.822	0.768	0.832

Neural networks perform slightly better than the other methods with most of the feature sets, but in general the results are quite disappointing, especially regarding the most important outcome, i.e. sensitivity, the power to correctly detect dropouts. Sensitivity, in fact, is almost null for the naïve Bayes algorithms that do not use video-lectures usage features, but also with the other methods it ranges from 0.303 (k-nearest neighbors, basic features) to 0.6 (for neural networks on basic features

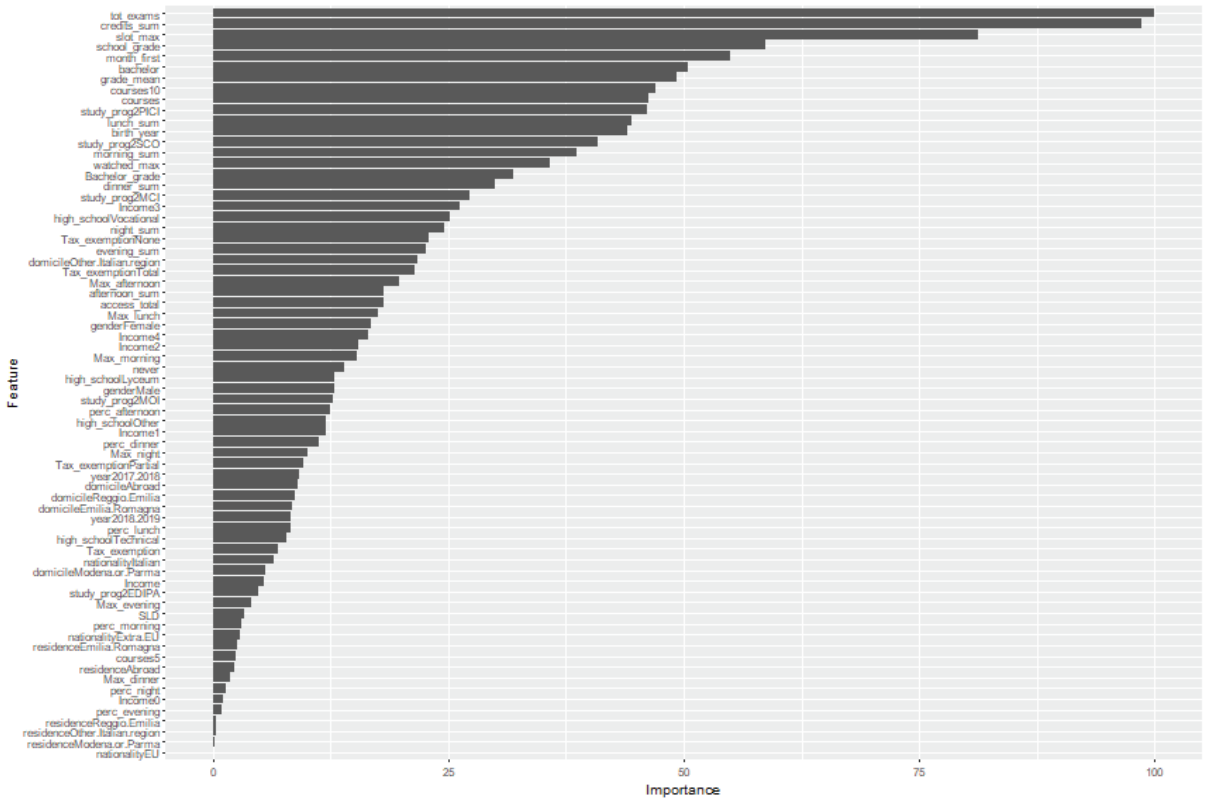
and academic achievements, and for naïve Bayes with all features). Moreover, accuracy is never high, while the somehow higher level of specificity can probably be explained by the slightly imbalanced structure of the data. Dropouts, in fact, are less than 30% of the observations, and consequently the trivial decision to always predict "no dropout" would guarantee a specificity larger than 0.7.

The comparison between performance measures with basic features and academic achievements, and performance measures with all features indicates that ONELab access data add very few (or nothing) to the usual features from administrative data. On the other side, in many cases the use of video-lectures usage only leads to the same performance as the use of the much larger and complex set of administrative data indicated as "basic features".

The most accurate and sensitive predictions (with the exception of the naïve Bayes) are those which include data on the academic achievements during the winter session of exams, even though this group of features only includes three variables, while the use of the large groups of basic features and ONELab usage features alone does not lead to outstanding results.

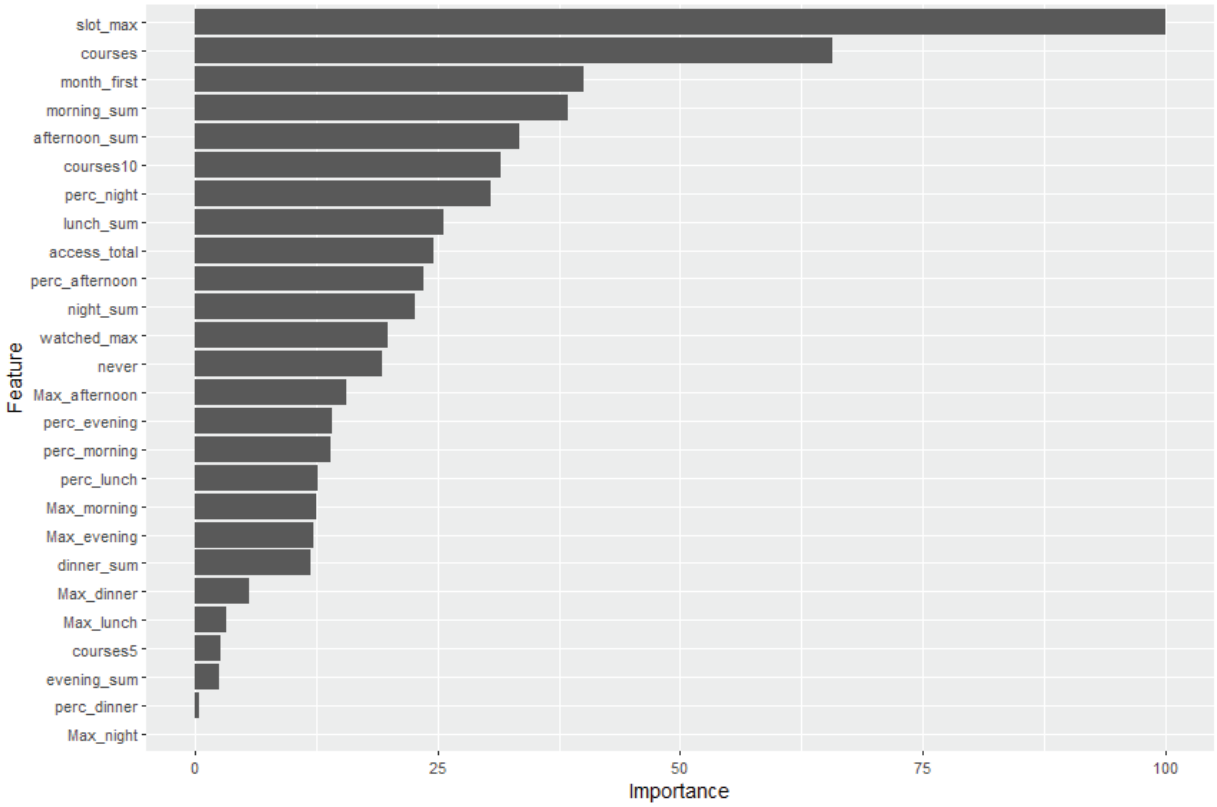
The importance of the academic achievement variables is confirmed if we look at the variable importance of the whole set of features with the neural networks method in Figure 1 (the variable importance obtained with the other algorithms is analogous); here we notice that the first positions are occupied by academic achievement variables ('tot_exams' and 'credits_sum'), followed by the number of video-lectures accessed in the favourite time slot ('slot_max'). The third feature on academic achievements is also among the ten most important variables.

Figure 1. Variable importance with neural network method; all features.



Limiting the variable importance analysis to the video-lectures usage features (Figure 2), the most important features are the number of lectures accessed in the favourite time slot ('slot_max'), and the number of different courses that have been accessed at least once ('courses'), which can probably capture the extent of time devoted to study. Some of these features, as well as the complete set discussed above, are partially redundant, and this impacts the differences in variable importance.

Figure 2. Variable importance with neural network method; video-lectures usage.



4. Conclusions

The general behaviour of the predictive models was in a sense disappointing, not especially with regard to the predictive power of the proposed variables on online access to lectures recordings, but for the overall scarce capacity to predict dropouts, even with traditional administrative and career data.

If compared to the results of similar studies (Del Bonifro et al., 2020; Berens et al., 2018) accuracy and sensibility are a bit lower. What is different in this study? First, in order to think of an early prediction, this study only considers the academic achievements obtained during the first exam session (January and February), while other studies include all the credits obtained in the first year and sometimes (for the prediction of dropouts occurring after the first year) also in the next years. In fact, although these analyses have not been shown here, data predictions based on the achievements of the whole academic year are a bit better but, at the time of the last exam session, a large amount of dropouts have already happened, arising the problem of a mutual dependency between dropouts and some features. On the other hand, when considering only enrollment or first semester data, also Berens et al. (2018) obtain accuracy levels lower than 0.7 and (in some cases) sensibilities under 0.5.

Moreover, in the attempt to reproduce the conditions in which a prediction predicts, in this study all the students who withdrew from university before June were excluded from the analyses. It's only 78 cases out of the original 2,342, but it would have raised the percentage of dropouts over 35% and, above all, these early dropouts were likely to be easier to detect (almost no exams, no credits, no access to the video-lectures), and they would have been about 10% of the total dropouts. But, again, we cannot call "prediction" the announcement of something that has already happened at the prediction time.

When adding the features on video-lectures fruition, the improvement was mostly negligible, and this is bad news. However, these data are costless, because they are automatically collected when students log in to access the lectures recordings, and harmless. Furthermore, the integration

of these features was useful to enhance the sensitivity in those classification models where sensitivity was alarmingly low. Even more promising: in most cases the predictions based only on ONELab access data are not worse than the predictions based on the whole set of features (except the academic achievement).

The early prediction issue remains at stake: is it actually possible to make an effective early prediction? Maybe a very early prediction, even before the winter session of exams, when students still have time to recover and learn how to study, or re-orientate to a different study programme? According to these results, and looking at the comparison to the results of other studies, when it comes to dropout predictions there is a trade-off between the need to detect students at risk as early as possible, and the growing inaccuracy of detection. Maybe, in the future perspective of a "University 4.0", students will be traced more steadily (monitoring their presence in the classroom, in the studying rooms, their use of online textbooks and exercises, and so on), and very early detection will come true. Or, who knows, it's mere science-fiction.

References

Agrusti, F., Bonavolontà, G., Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-Learning and Knowledge Society*, **15**(3), pp. 161-182.

Aina, C., Baici, E., Casalone, G., Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, **79**, pp. 1-16.

Berens, J., Schneider, K., Görtz, S. Oster, S., Burghoff, J. (2018), Early detection of students at risk – Predicting student dropouts using administrative student data and machine learning methods. CESifo Working Paper No. 7259, CESifo, Munich.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), pp. 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, (CA).

Cover, T.M., Hart, P.E. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**(1), pp. 21-27.

Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P. (2020). Student dropout prediction, in I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, E. Millán (eds). *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science*, vol 12163. Springer, Cham.

Domingos, P., Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, **29**, pp. 103-130.

Furini, M., Galli, G., Martini, M.C. (2020). An online education system to produce and distribute video lectures. *Mobile Networks and Applications*, **25**, pp. 969-976.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, **28**(5), pp. 1–26.

OECD (2022). *Education at a Glance 2022: OECD Indicators*. OECD Publishing, Paris.

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.

doi.org/10.26398/asaproc.00121

Statistical indicators for the analysis of social sustainability and human well-being

Antonella Massari[a], Paola Perchinunno[a], Samuela L'Abbate[a], Domenico Leogrande[a], Lucia Mongelli[b]
[a] Department of Economics, Management and Business Law - University of Bari (Italy)
[b] National Institute of Statistics (ISTAT), Bari, (Italy)

1. Introduction

A fundamental prerequisite for a correct statistical analysis of social sustainability is the need to share a univocal definition of the concept itself. Social sustainability, as defined by the United Nations, is 'the process of creating sustainable environments that promote physical and mental well-being by addressing the needs of individuals in the places where they live and work'.

The objective of this work is to identify multidimensional statistical methodologies capable of measuring social sustainability in its complexity and being of support to political and economic planning for targeted interventions in the territory. In this study the relationships between the indicators of social sustainability and human well-being are analyzed related to Italian provinces. The model used considers the indicators available in the report on fair and sustainable well-being (BES). Two different multidimensional statistical analysis methodology will be used: totally fuzzy and relative method and a spatial clustering method (density based spatial clustering of application with noise).

The first method uses the technique of fuzzy sets to obtain a measure of territorial and tourism sustainability incidence within a population from the statistical information provided by a plurality of indicators (totally fuzzy and relative method*)*. The second is a method of clustering territorial units to obtain areas of high (or low) intensity of the phenomenon which allow the aggregation of contiguous spatial units homogeneous in relation to the phenomenon under study (DBSCAN method*)*.

2. Approaches and methodologies for social sustainability analysis

2.1 The fuzzy approach

The development of fuzzy theory initially stems from the work of Zadeh (1965) and subsequently was conducted by Dubois and Prade (1980). The fuzzy theory develops starting from the assumption that each unit is not univocally associated with only one but simultaneously with all the categories identified based on links of different intensity (degrees of association). The first measurement based on the fuzzy set theory, named TF (totally fuzzy), as suggested by Cerioli and Zani (1990). This logic can be applied to both continuous and ordinal variable cases. However, in the latter case, the maximum and minimum values can be determined by assuming the value of the lowest category as minimum and the highest as maximum.

Cheli and Lemmi (1995) have proposed a generalization of this approach, called totally fuzzy and relative (TFR). This method is also called "totally relative" because the value of the membership function is entirely determined by the relative position of the individual in the distribution of the population. The fuzzy TFR approach consists in defining the measurement of an individual's degree of belonging to the totality fuzzy, included in the interval between 0

(with an individual who does not demonstrate a clear belonging) and 1 (with an individual who demonstrates a clear belonging).

If we suppose to observe k indicators for each family, the function of belonging of the i-th family to the blurred subset can be defined as follows:

$$f(x_{i.}) = \frac{\sum_{j=1}^{k} g(x_{ij}).w_j}{\sum_{j=1}^{k} w_j} \qquad i = 1,.....,n \tag{1}$$

where $g(x_{ij})$ measures the probability of membership of each unit and w_j derives from a weighting system, as given by generalizing the one proposed by Cerioli and Zani (1990):

$$w_j = \ln\left[1/\overline{g(x_j)}\right], \qquad \text{where} \quad \overline{g(x_j)} = \frac{\sum_{i=1}^{n} g(x_{ij})}{n} \qquad (j = 1,2,...,k). \tag{2}$$

2.2 The DBSCAN approach

The fuzzy approach responds to the need to identify spatial areas and/or population subgroups characterized by situations of hardship or high social exclusion because it allows a measure to be defined of the degree to which they belong to the fuzzy set of the deprived. It is possible to employ methods of clustering territorial units to obtain areas of high (or low) intensity of the phenomenon by using clustering methods that allow the aggregation of contiguous spatial units homogeneous with respect to the phenomenon under study. DBSCAN (density based spatial clustering of application with noise), Ester et al. (1996), was the first clustering method based on spatial density. The basic idea is to define a new cluster or extend an existing cluster, the surroundings of a point of a given radius ε, must contain at least a minimum number of MinPts, i.e., the density of the points must exceed a certain threshold. DBSCAN as input needs the parameters ε and MinPts to be fixed. Their choice is crucial since they determine whether a group of points is a cluster or simply noise. Let d denote the distance of a point p from its k -th nearest neighbour; within a distance d from p are then contained k+1 points (barring special cases in which multiple points are at the same distance d from p). Having fixed k, we define a k -dist function, from the dataset D to the set of real numbers, that assigns to each point in D the distance to its k -th nearest neighbour. Then the points in the dataset are sorted in descending order with respect to k -dist, resulting in a graph called the sorted k -dist graph. This graph provides information about the density distribution of the dataset: in particular, by choosing an arbitrary point p and fixing MinPts to k and ε to k -dist(p), all points having equal or lesser k -dist will be core points. Having chosen a threshold point on the graph, to which corresponds a value of k -dist and hence ε, all patterns preceding it turn out to be noise while those following it will be assigned to clusters. The authors recommend choosing as a threshold the point at the first re-entrant that can be seen in the graph's trend. Such determination of ε has ambiguities within it since the graph does not have a sharp dip but often fluctuations are evident so the choice of ε is reduced to a trial-and-error choice of the most appropriate one.

3. Data set and indicators

As the basis of analysis for the study of social deprivation present in the territory at present, attention has been focused on the "Equitable and Sustainable Welfare - Bes" data ISTAT Rapporto Bes (2022), ed. 2023: releases the annual update of the territories' Bes indicators, referring to the 107 Italian provinces and metropolitan cities, and the web-based database query

system. The 2023 edition contains 70 indicators of well-being at the provincial level (Nuts3), in historical series and complete with all comparison values for higher territorial levels. The available data provide an understanding of the social, economic, demographic, and environmental conditions of the regional territory, along with measures that reflect the levels achieved in terms of community well-being. Specifically, 3 sets of indicators were considered at the provincial level, each containing within its multiple indicators for each of the areas of distress identified relating to: Demographic aspects; Employment aspects; Personal safety aspects.

The indices were chosen to identify different levels of hardship and were calculated so that high index values correspond to high levels of difficulty.

The TFR method was applied on the data of all the Italian province obtaining a value of the individual w_i weights, which varies according to the level of importance in determining the degree of quality of the situation. For each set of indicators, the fuzzy value and its related values were calculated. High values are significant of situations of hardship, conversely low values are significant of situations of no hardship.

The provinces were aggregated through the DBSCAN algorithm, based on the fuzzy value or not. The provinces are represented in maps whit white colours for the provinces that were classified by the algorithm as noise points, around which there are fewer provinces than MinPts and none of them is a core point. That is, the noise provinces are at a distance greater than eps from any core point province. In the same figures, we show graphs of the k-distances for each of the three sets of indicators, and the optimal value empirically chosen to be assigned to ε is highlighted. Increasing even slightly the value of ε concentrates the points into a single cluster, while decreasing the value of ε increases the noise points and increases the clusters.

As regards the demographic aspects, the following three indicators were considered: Life expectancy (inverse); Avoidable mortality (0-74 years); Persons who have not completed high school (25-64).

By first applying the fuzzy method to synthesize the data of the three indicators and then the DBSCAN method on the basis of fuzzy value, six clusters are obtained. The legends show the respective fuzzy mean values for each cluster obtained, the darker the colour the greater the distress for the provinces aggregated in the cluster.

Six clusters were obtained; characterized by medium-high fuzzy values formed by the large provinces of Piedmont and Lombardy combined to those located in the southern part of Sardinia, and those in Campania; there is not a significant demographic inequality in the clusters formed by the provinces of Veneto and by by a part of the other provinces of Emilia-Romagna and Tuscany.

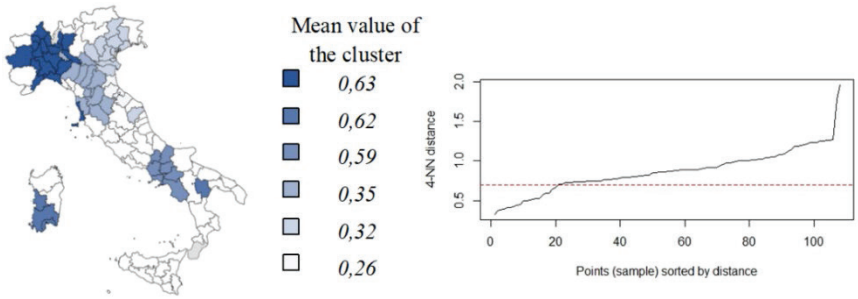


Figure 1. Demographic hardship with fuzzy mean value, graph of the k-distances (ε=0.7 MinPts=4).

In contrast, directly applying the DBSCAN method to synthesize the three indicators yields six clusters, fewer noise points, and thus more provinces in central Italy and the provinces of

Sicily and Calabria are also considered as clusters.

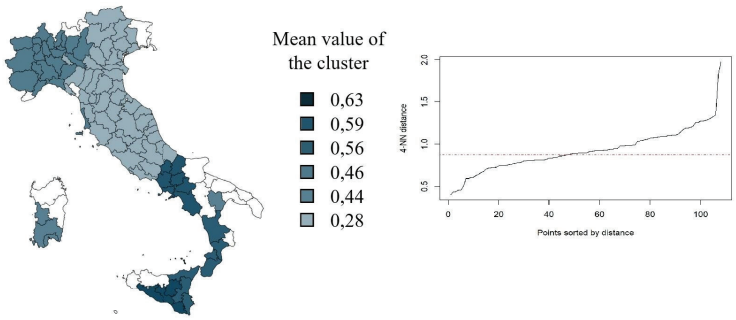


Figure 1 bis. Demographic hardship with indicators means, graph of the k-distances (ε=0.87 MinPts=4).

As regards the demographic aspects, the following three indicators were considered: labor non-participation rate; youth (15-29 years old) labor non-participation rate; pensioners with low pension income.

By first applying the fuzzy method to synthesize the data of the three indicators and then the DBSCAN method on the basis of fuzzy value, six clusters are obtained. Among them, one emerges, comprising the provinces of Campania with a high fuzzy value, indicative of significant occupational inequality. Two other clusters show a medium-high fuzzy value: one includes the provinces of Sardinia, and the other comprises the provinces of Central Italy.

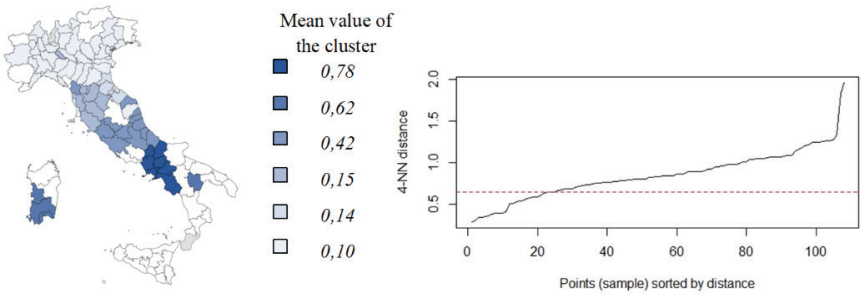


Figure 2. Employment hardship with fuzzy mean value, graph of the k-distances (ε=0.64, MinPts=4).

In this case, directly applying the DBSCAN method for the synthesis of the three indicators yields six clusters, and the situation appears to be unchanged.

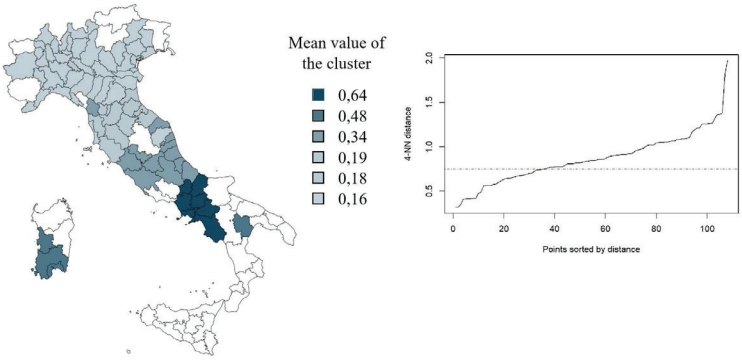


Figure 2 bis. Employment hardship with indicators means, graph of the k-distances (ε=0.745 and MinPts=4).

3.4 Indicators of personal safety hardship

As regards the personal safety aspects, the following four indicators were considered: correctional institution crowding; other reported fatal crimes; reports of residential theft; claims of pickpocketing.

By first applying the fuzzy method to synthesize the data of the three indicators and then the DBSCAN method on the basis of fuzzy value, seven clusters are obtained. The cluster composed of the provinces of La Spezia, Mantua, Modena, Parma, and Reggio shows an average fuzzy value, along with the cluster formed by the provinces of Lombardy, Veneto, and Liguria, and the one consisting of some provinces of Tuscany.

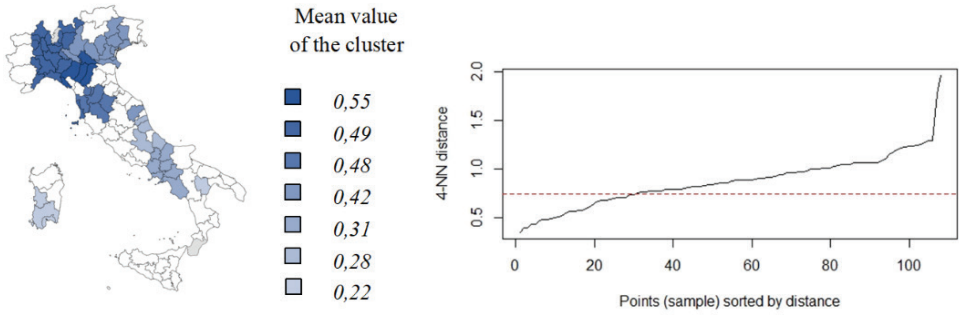


Figure 3. Personal safety hardship with fuzzy mean value, graph of the k-distances (ε=0.75, MinPts=4).

Finally, directly applying the DBSCAN method for the synthesis of the three indicators results in five clusters, more provinces are included in the clusters, so the number of noise points is decreased, and some provinces of Calabria and Sicily are also considered as clusters.

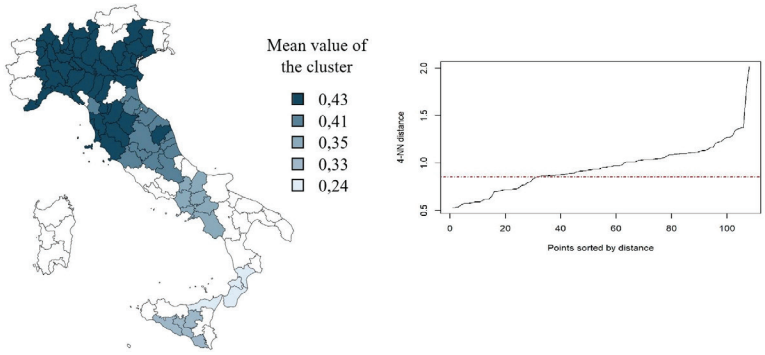


Figure 3 bis. Personal safety hardship with indicators means, graph of the k-distances (ε=0.855, MinPts=4).

4. Conclusions and future development

A first idea emerging from the present work is related to the ability to describe territorial phenomena through an integrated model, which starts from the construction of social indicators, multidimensional in nature, and then adopts models capable of identifying areas at risk of hardship, Perchinunno et al. (2023). The application of the Totally Fuzzy and Relative method allows us to obtain for each province a "fuzzy" level of inequality between zero and one, given by the synthesis of the values of the indicators of each set of inequalities. The advantage of DBSCAN lies in the algorithm's ability to identify clusters of arbitrary shapes and bring together spatially close areas, in this case provinces, with the same intensity of the phenomenon considered i.e. the fuzzy value or the indicators means.

The DBSCAN algorithm applied was the same for both cases examined with the difference of having used as input a data matrix with the spatial coordinates and a single fuzzy summary value in the first case, and in the second case in addition to the spatial coordinates also the indicator values for each set. By increasing the size of the dataset in the second case, DBSCAN increases the execution time, however, for the benefit of greater accuracy, allowing for greater scalability and adaptability of the clustering method in high dimensionality spaces, at the expense of a quantitative error threshold that can be made arbitrarily small.

An initial descriptive analysis shows that the levels of discomfort are not clearly localized to southern Italy but are scattered throughout the country in a diverse manner.

This implies the need for more targeted economic efforts aimed at supporting regional policies to create levels of social sustainability in the territory.

A more recently developed algorithm, Campello et al. (2013), built upon DBSCAN is HDBSCAN, which, unlike DBSCAN, can identify noise and clusters of varying density. Unlike DBSCAN, HDBSCAN requires only one input parameter, MinPts. DBSCAN's additional eps parameter is often considered a very unintuitive parameter to tune. By contrast, MinPts for HDBSCAN feels a lot more intuitive to set. Finally, HDBSCAN is faster and computationally more efficient than DBSCAN. As a future development we can think of comparing the two algorithms and evaluate to get a better classification.

References

Campello et al. (2013). Density-based clustering based on hierarchical density estimates. PAKDD 2013. *Advances in Knowledge Discovery and Data Mining*, pp. 160–172.

Cerioli, A., Zani, S. (1990). A fuzzy approach to the measurement of poverty. In C. Dagum, M.Zenga (eds.) *Income and Wealth Distribution, Inequality and Poverty*. Springer Verlag, Berlin. O

Cheli, B., Lemmi, A. A (1995). Totally fuzzy and relative approach to the multidimensional analysis of poverty. *Economic Notes* 24, (1), pp.115-134.

Dubois, D., Prade, H. (1980). *Fuzzy Sets and Systems*. Academic Press, Boston, New York, London

Ester M, Kriegel H.-P, Sander J., Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, Portland.

Istat (2022). Il Benessere Equo e Sostenibile in Italia. Data di pubblicazione: 20 aprile 2023 https://www.istat.it/it/files//2023/04/Bes-2022.pdf

Montrone, S., Perchinunno, P., Rotondo, F., Torre, C. M., Di Giuro, A. (2009). Identification of hot spots of social and housing difficulty in urban areas: Scan statistic for housing market and urban planning policies. In: B. Murgante, G. Borruso, A. Lapucci (eds.) *Geocomputation and Urban Planning, Studies in Computational Intelligence*, 176, Springer Verlag Berlin Heidelberg, pp. 57-78.

Perchinunno P., Rotondo F., Torre C. M. (2008). A multivariate fuzzy analysis for the regeneration of urban poverty. In: M. Gervasi, T. Laganà, G. Mun, (eds.) *ICCSA 2008, Part I. LNCS,* 5072, Springer, Heidelberg, pp. 137-152

Perchinunno P., Massari A., L'Abbate S., Carbonara M. (2023). A spatial statistical approach for the analysis of urban poverty. In *Lecture Notes in Computer Science* - ISBN: 978-3-031-37110-3 vol.14106.

Zadeh, L.A (1965). *Fuzzy Sets, Information and Control*. 8(3), pp. 338-353.

doi.org/10.26398/asaproc.00122

Health and discrimination among refugees in Germany

Daria Mendola[a], Annalisa Busetta[b]
[a] Department of Psychology, Educational Science and Human Movement, University of Palermo, Palermo, Italy.
[b] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy.

1. Introduction

Refugee health is not only a matter of socioeconomic conditions but is also highly dependent on perceptions and feelings. Research on refugees has traditionally concentrated on trauma and its effects on health, as well as the impact of socioeconomic conditions in host countries on refugees' well-being. In this paper, we specifically investigate the multiple determinants of self-assessed health among refugees who have recently arrived in Germany, exploring how health conditions relate to perceived discrimination, considering also loneliness, life satisfaction, and the role of individual resilience in mitigating these effects. Since health, life satisfaction, resilience, and perception of discrimination can be considered latent constructs, we estimated a structural equation model to explore relationships among them.

Recently, scholars have begun examining refugees' subjective evaluations of their life situations, including factors like discrimination, loneliness, social isolation, and life satisfaction (Ambrosetti et al., 2021; Mendola et al., 2024). Discrimination is a major stressor for immigrant health, linked to poor mental, physical, and behavioural outcomes, including anxiety, depression, cardiovascular issues, and unhealthy behaviours (Williams et al., 2003). The subjective experience of discrimination is particularly impactful, triggering stress responses that affect both psychological and physiological well-being (Andrighetto et al., 2013; Ziersch et al., 2020).

Perceived discrimination is widespread among immigrants and ethnic minorities in both the U.S. and Europe, though its nature varies by group, often tied to ethnic origin and skin colour (Straiton et al., 2019; Andriessen et al., 2014). Studies highlight its role in limiting opportunities for refugees and hindering access to healthcare (Szaflarski and Bauldry, 2019).

2. Data and methods

- Sample

Our analyses rely on first-time respondents of the first two waves of the IAB-BAMF-SOEP survey of refugees v34 (Brücker et al., 2021) conducted in Germany (i.e., samples M3-M4 and M5). The target population comprises asylum seekers and refugees who entered Germany between 2013 and January 2016 and applied for asylum or temporary protection in 2013-2017, regardless of the application results. In addition, the respective household members (aged 18 or older) were also interviewed. The participants were selected from the Central Register of Foreigners (AZR) using a two-stage, disproportionate stratified sampling design where the primary sampling units (PSUs) are the regional clusters of immigration offices and the second stage units are the individuals selected on a disproportional sampling scheme that ensured minimum sample sizes for subgroups of refugees (more information on the sampling in Kühne et al. 2019).

The survey gathers information on individual socio-demographic characteristics, including education, labour skills, current work situation, language skills, social participation, feelings, perceptions, attitudes, various psychometric scales, and household- and family-level information. Our "working" sample included 6,054 adults with around 70% of four main nationalities: Afghan, Eritrean, Iraqi and Syrian. The mean age ranges from 27 years among Eritreans to 32 years among Syrians and the percentage of women in the community ranges from 25% of Eritreans to 32% of

Syrians.

- Health and other related constructs

Our main interest here is to determine the factors associated with the current health status of refugees living in Germany, with a particular interest in discrimination. Since health is a multidimensional concept that includes physical and mental well-being, it is difficult to have a consistent measure valid across populations and subgroups (WHO, 2018). In this paper, the health conditions of refugees were measured by three observed indicators: a) satisfaction with current health status, b) self-reported pain, and c) a measure of psychological health. Particularly:

a) satisfaction with current health status is considered a good indicator of "true" individual health status that provides a useful summary of how people perceive their overall health status. Particularly, the study by Bauer et al. (2020) on refugees in Germany showed that health satisfaction is also highly correlated with self-assessed health and mental health. Satisfaction with current health status is measured by the question "How satisfied are you today with your health?" with responses on a scale from 0 "totally dissatisfied" to 10 "totally satisfied".

b) physical pain is measured by the question "How often in the last four weeks did you suffer from severe physical pain?". This question is on a five-point Likert scale from "always" to "never".

c) presence of depression in the last four weeks according to the question, "How often in the last four weeks did you feel in low spirits and melancholy?" As the previous one, this question is on a five-point Likert scale from "always" to "never".

To test the complex model of relationships around health, we also included in our analysis refugees' perceptions of discrimination[1] and loneliness[2] and their life satisfaction[3] and resilience[4], defined and measured considering only the questions formulated in the same way in the two waves considered.

- The structural equation model

Health, resilience, life satisfaction, perceived discrimination and loneliness are abstract and multidimensional concepts that can be impacted by inaccuracies or deviations in observed variables, arising from random or systematic factors. To address potential measurement errors, we chose to represent these constructs as latent variables and employed a SEM (structural equation model, Kline, 2023). The model formalises relationships among latent variables and statistically tests for the validity of the theoretical model; it provides estimates of the direct and indirect effects among latent variables. A SEM comprises the "measurement model" and the "structural model". The main interest usually resides in the last one, which sets causal relationships and associations among latent variables.

Here, our main interest is evaluating the role that the perception of discrimination plays in influencing the health of refugees directly and/or indirectly, controlling for other psychosocial factors (i.e., resilience, loneliness and life satisfaction). The paper's aim is to test the direct effects of these four latent variables on health, and, at the same time the indirect effects of perceived

[1] Perceived discrimination is defined as the perception of unfair treatment based on characteristics such as gender, age, ethnicity, and sexual orientation. It may differ from legal or scientific definitions of discrimination. Our study assesses perceived discrimination through refugees' sense of belonging in Germany, experiences of ethnic discrimination, and concerns about anti-foreigner sentiment.

[2] Loneliness is defined as the experience of lacking quality social contact, specifically emotional loneliness. The study measures loneliness using three items from the UCLA Loneliness Scale, focusing on how often individuals miss the company of others, feel socially isolated, and miss people from their country of origin.

[3] Life satisfaction is defined as the extent to which a person finds life rich, meaningful, and of high quality. In this study, life satisfaction is measured through satisfaction with current living arrangements, current health, and overall life satisfaction.

[4] Resilience is the ability to successfully adapt to challenging experiences, demonstrating inner strength, flexibility, and effective coping. The study measures resilience using the Brief Resilient Coping Scale, focusing on the ability to think of solutions to difficult situations, control reactions, develop from challenges, and balance losses.

discrimination on health through the mediator variables loneliness and life satisfaction. Control variables (namely, age and sex) are considered to have a direct effect on health, while nationality (five categories: Syrian, Afghan, Eritrean, Iraqi and the rest of the world) has a direct effect on the latent variable perceived discrimination.

3. Results

The analysis of the three items on health (their satisfaction with general health conditions, physical pain and the presence of anxiety/depression in the four weeks before the interview) provides a first portrait of refugees' health conditions in our sample. The general portrait reveals that the average satisfaction with the health of refugees is 7.9 (from 0 to 10); furthermore, 14.4% reported severe pain in the last 4 four weeks, and approximately 30% reported anxiety or depression in the last four weeks.

Regarding discrimination, the perception of xenophobia against foreigners inside German society is quite strong, with around 26% of refugees declaring being very or somewhat concerned about it. However, only 5% were discriminated against because of their origin, and around 4% felt barely or not at all welcome in Germany at the moment of the interview.

Interestingly, in this framework, not all refugee groups experience the same health and discrimination conditions. This is because, frequently, different treatments are reserved for different refugee populations acknowledged as "deserving" and "not deserving" to be hosted and supported by societies and welfare states (Edge and Newbold, 2013; Andriessen et al. 2014). Our study highlighted how, in Germany, Eritreans have a lower incidence of bad health conditions, report higher satisfaction with their health, and lower incidence of physical pain and depression symptoms (see Figure 1). Noteworthy, subjective health indicators, such as self-reported pain and satisfaction with health, may be affected by reporting bias related to individual perception and differences in cultural traits. Hence, further investigation, beyond the scope of this study and not possible with the current survey, would be necessary to disentangle real health conditions and reported ones.

More nuanced is the situation of experienced and perceived discrimination of refugees living in Germany. Figure 2 shows how xenophobia is significantly less perceived among Syrians, Eritreans and Iraqis (when compared to other ethnic groups), while no differences among refugees can be statistically assessed about experienced episodes of discrimination due to ethnic origin (whose incidence is generally low and around 5% in all the groups here considered). Finally, Syrians perceived themselves as welcome in Germany more than Afghans and the residual group ("others") do.

Focussing on the results from the structural equation model (Figure 3), it emerged that all the relationships among the latent variables are highly significant and have the expected direction. The goodness of fit indexes (Hu and Bentler, 1999) are more than satisfactory (CFI and TLI are equal to 0.931 and 0.955, respectively; the RMSEA is equal to 0.047). Higher loneliness and discrimination are associated with a lower health of refugees, while higher resilience and life satisfaction are associated with better health. Particularly, feeling or being discriminated against is negatively related to refugees' health directly and via the positive relations between discrimination and loneliness. Perceived and experienced discrimination also negatively connects to refugees' life satisfaction.

Furthermore, lonely refugees experience lower life satisfaction and, consequently, lower health. Note that the cross-sectional data selection on which this paper is based did not allow for the assessment of the real direction of some links, which could also be conceived differently. Furthermore, some relations can be conceived in both directions with valid literature arguments. For this reason, any of the relations in Figure 3 should be carefully interpreted as associations, while causal effects should need a different analysis strategy.

Figure 1 - *Health conditions of refugees by country of origin (confidence interval 95%)*

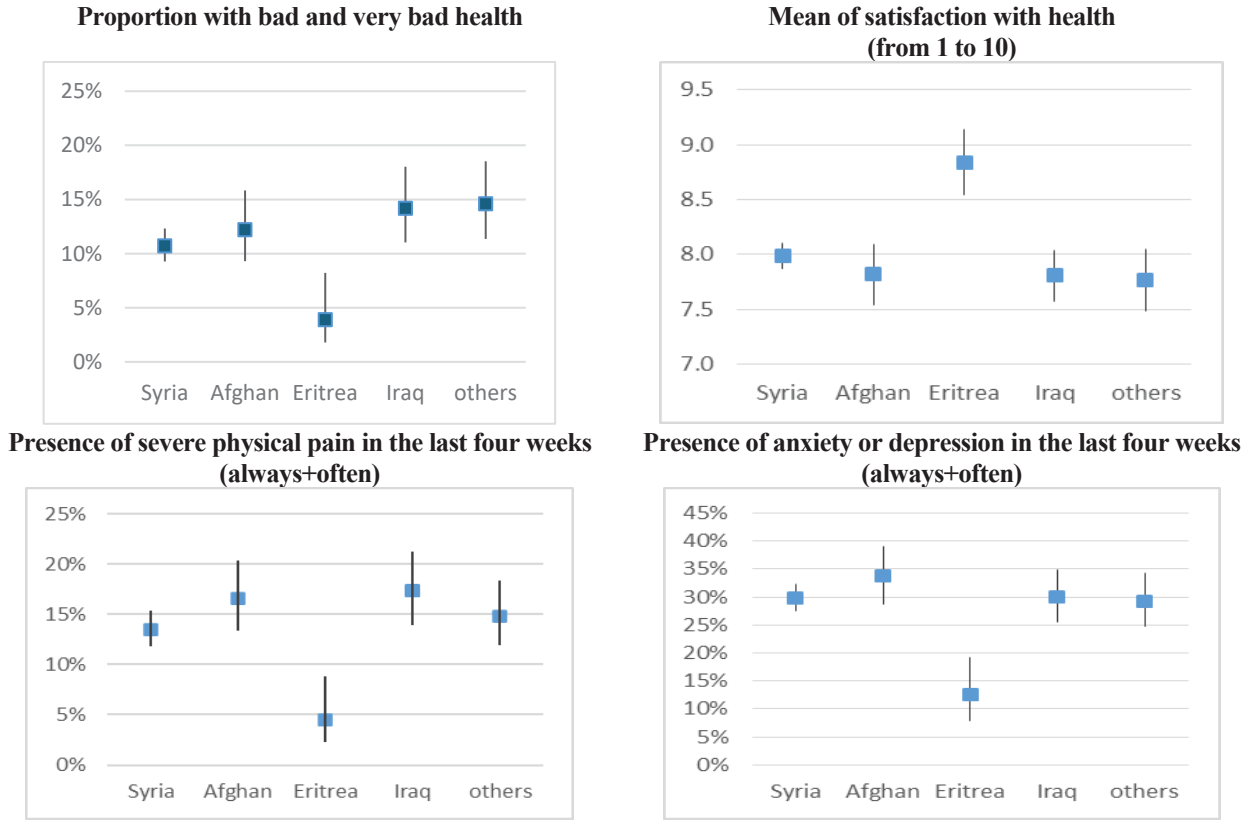


Figure 2 – *Discrimination among refugees by country of origin (confidence interval 95%)*

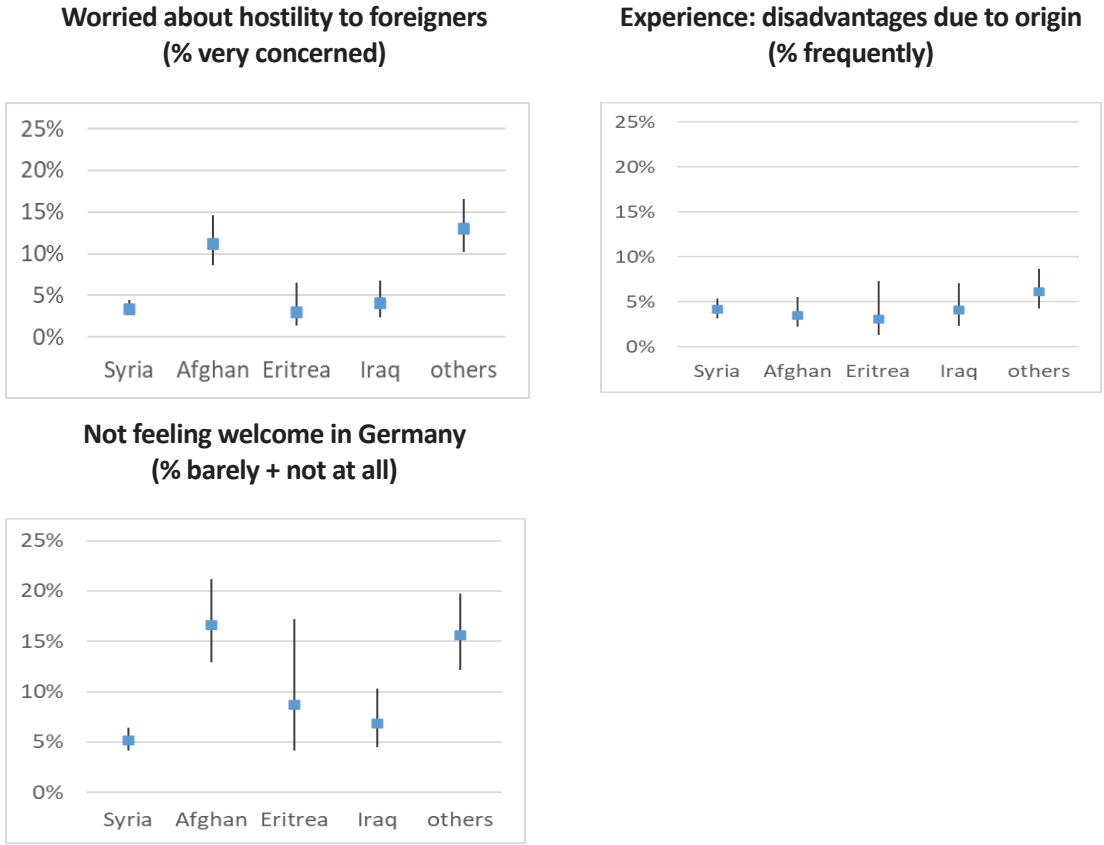
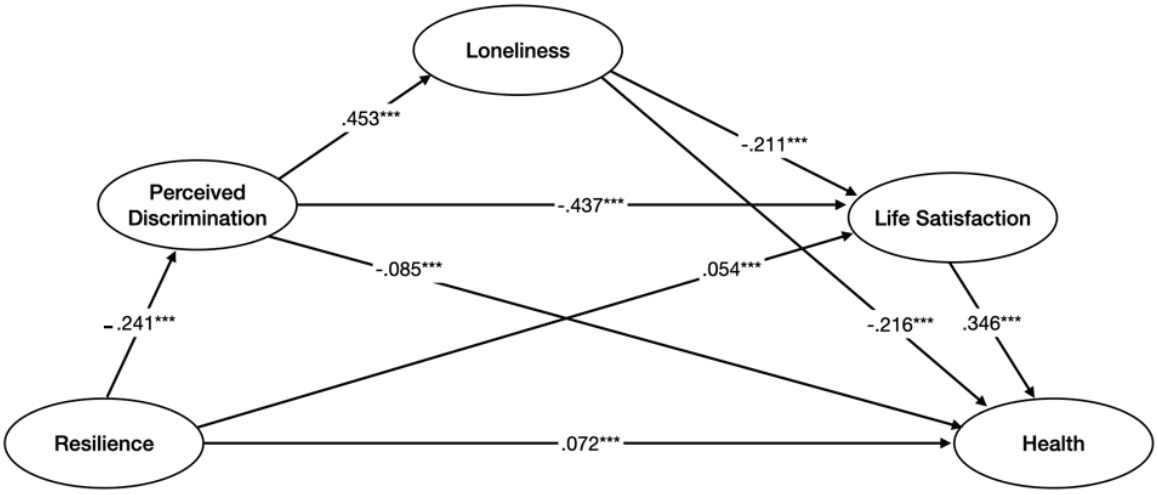


Figure 3 – *The structural equation model (path coefficients and significance levels)*



Note: p-value symbols: *** <0.01; ** <0.05 * <0.10

4. Conclusions

The German welfare state provides material assistance to refugees within the official reception system. However, the integration process remains challenging for migrants who were forced to leave their home countries. Refugees often face harsher living conditions, poorer economic integration, and deeper feelings of loneliness compared to other migrants. Additionally, some refugees experience greater discrimination, primarily due to their nationality and the prevailing media and political discourse.

This study focuses on the significance of individual perceptions and feelings about health, deliberately setting aside well-known objective health determinants (e.g., living conditions, income, accommodation, nutrition, etc.). We intentionally centred our analyses on the voices of refugees, avoiding the imposition of objective indicators (when available) in the survey questionnaire.

In this context, resilience is negatively associated with the perception of discrimination. Although resilience is an intrinsic characteristic of individuals, social policies can enhance it by, e.g., promoting more inclusive initiatives that actively involve refugees in their integration. As research has demonstrated, discrimination has a significant negative impact on reported health—both directly and indirectly—by exacerbating loneliness and reducing life satisfaction. Strengthening refugees' sense of belonging to the host society can mitigate the perception of discrimination and loneliness. Additionally, promoting joint initiatives involving German natives and refugees can help reduce anti-foreigner sentiment.

This study is not without limitations. First, it is grounded on cross-sectional data, which severely limits the assessment of causal links. Noticeably, a necessary condition to include a longitudinal structure in an SEM model is the availability of at least three time-points for the interview. Unfortunately, the IAB-BAMF-SOEP survey of refugees, records the health variables only every second year. These two conditions, together with the presence of missing data in many variables, enhance the attrition in data selection so that the number of refugees suitable for a causal model was too small.

Second, as previous descriptive analyses showed, people with different nationalities can report different health quality and perceived and experienced discrimination, due to objective and subjective factors and cultural differences. Although this might be an indication in favour of a multigroup structural equation modelling, the potential confounding effect of pooling all nationalities together is not independent of the different cultural norms in reporting health problems. Exploring the group-specific variation is beyond the aim of this paper, although we acknowledge

this could hide potential differences in health outcomes and perceptions across refugee groups.

References

Ambrosetti, E., Dietrich, H., Kosyakova, Y., Patzina, A. (2021). The impact of pre-and postarrival mechanisms on self-rated health and life satisfaction among refugees in Germany. *Frontiers in Sociology*, ***6***, 693518.

Andriessen I, Fernee H, Wittebrood K. (2014). *Perceived Discrimination in the Netherlands*. The Hague: The Netherlands Institute for Social Research. Available from: https://www.researchgate.net/publication/265412943_Perceived_discrimination_in_the_Netherlands.

Andrighetto, L., Durante, F., Lugani, F., Volpato, C., Mirisola, A. (2013). Obstacles to intergroup contact: When outgroup partner's anxiety meets perceived ethnic discrimination. *British Journal of Social Psychology*, **52**(4), pp. 781-792.

Bauer, J.M., Brand, T., Zeeb, H. (2020) Pre-migration socioeconomic status and post-migration health satisfaction among Syrian refugees in Germany: A cross-sectional analysis. *PLoS Med*, **17**(3), e1003093.

Brücker, H., Liebig, S., Leistner-Rocca, R., Goebel, J., Grabka, M.M., Rother, N., et al. (2021). *IAB-BAMF-SOEP Survey of Refugees – version 1619 v1*. Research Data Centre of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB). DOI: 10.5684/soep.iab-bamf-soep-mig.2019.

Edge, S., Newbold, B. (2013). Discrimination and the health of immigrants and refugees: Exploring Canada's evidence base and directions for future research in newcomer receiving countries. *Journal of Immigrant and Minority Health*, **15**, pp. 141–148.

Hu, L.T., Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, **6**(1), pp. 1-55.

Kline, R.B. (2023). *Principles and Practice of Structural Equation Modelling*. Guilford publications, New York.

Kühne, S., Jacobsen, J., Kroh, M. (2019). Sampling in times of high immigration: the survey process of the IAB-BAMF-SOEP survey of refugees. *Survey Methods: Insights from the Field.* ISSN 2296-4754, FORS (Swiss Foundation for Research in Social Sciences), Lausanne, https://doi.org/10.13094/SMIF-2019-00005

Mendola, D., Parroco, A.M., Li Donni, P. (2024). We made it to Germany… and now? Interdependent risks of vulnerability for refugees in a high-income country. *Journal of Ethnic and Migration Studies*, ***50***(4), pp. 1059-1079.

Straiton, M. L., Aambø, A. K., Johansen, R. (2019). Perceived discrimination, health and mental health among immigrants in Norway: The role of moderating factors. *BMC Public Health*, **19**(1), pp. 1-13.

Szaflarski, M., Bauldry, S. (2019). The effects of perceived discrimination on immigrant and refugee physical and mental health. In *Immigration and Health* (pp. 173-204). Emerald Publishing Limited.

Williams, D.R., Neighbors, H.W., Jackson, J.S. (2003). Racial/ethnic discrimination and health: Findings from community studies. *American Journal of Public Health*, **93(2)**, pp. 200–208.

World Health Organization - WHO. (2018). *Report on the Health of Refugees and Migrants in the WHO European Region: No Public Health without Refugee and Migrant Health*. Copenhagen: WHO. Regional Office for Europe. ISBN 978-92-890-5384-6.

Ziersch, A., Due, C., Walsh, M. (2020). Discrimination: A health hazard for people from refugee and asylum-seeking backgrounds resettled in Australia. *BMC Public Health*, **20(1)**, pp. 1-14.

doi.org/10.26398/asaproc.00123

Estimating overdispersion in non-transformed scores of ANVUR's department ranking procedure: A Beto-multinomial model

Giorgio E. Montanari[a], Marco Doretti[b]
[a] Department of Political Science, University of Perugia, Italy.
[b] Department of Statistics, Computer Science, and Applications, University of Florence, Italy.

1. Introduction

As it is well-known, the National Agency for the Evaluation of Universities and Research Institutes (In Italian, ANVUR) is in charge of the periodic production of an academic departments' ranking, which is used to orient governmental funding policies. Such a ranking procedure relies on a dedicated performance index, known as the Standardized Index of Department Performance (in Italian, *Indice Standardizzato di Performance Dipartimentale*, ISPD), which is computed for each department as a function of the standardized scores attributed by an expert committee to a selected set of department's scientific papers.

In short, the ANVUR approach transforms a department-level performance indicator y via the cumulative density function (CDF) of a standard normal distribution (ANVUR, 2022), here denoted by Φ as usual. For each department, the resulting value $\Phi(y)$ can be then interpreted as the probability that a randomly selected department performs no better than it, with possible differences in their staff composition being accounted for. Thus, higher probabilities correspond to better performances. The whole framework assumes that the values of y (i.e., before Φ transforming) can be regarded as independent draws from a standard normal distribution. To obtain the ISPD values, the above probabilities are then rounded in a way such that - when expressed as a percentage - they correspond to either integer or semi-integer numbers (e.g., 82% or 77.5%).

Under this scheme and assumptions, the distribution of the department values of these rounded probabilities should be approximately uniform along the sequence [0, 0.005, 0.010, …, 0.990, 0.995, 1.000], that is, across the "rounding points" of the unit interval. However, empirical evidence from the 2017 ANVUR exercise showed a U-shaped distribution for the rounded $\Phi(y)$ values (ROARS, 2022), with symmetric peaks in correspondence of the interval boundaries, i.e., on 0 and 1. Such a distribution is rather far from the uniform distribution, and it suggests that the standard deviation σ of the underlying probability distribution of y is indeed greater than 1.

This increased variability is most likely due to the presence of un-modelled positive correlation within academic departments, which has three main causes: i) each scholar typically provides more than one paper, ii) papers are often co-authored by scholars affiliated to the same department, and iii) departments typically tend to hire scholars with similar scientific capabilities. Importantly, this intra-departmental correlation affects the overall validity of ISPD (Montanari and Doretti, 2023), and adjusting approaches based on knowledge of σ should be put forward to build a fairer performance indicator.

In this setting, the crucial task of estimating σ is challenged by the fact that only the rounded values of $\Phi(y)$ are observed, resulting in a loss of information. In particular, for units with extreme values (0 and 1), it is not possible to revert to the normal scale, since the inverse transformation Φ^{-1} does not return finite numbers. Such a problem is particularly relevant for the ANVUR data, where - as mentioned before - the portion of departments that end up taking these extreme values is rather large.

A very preliminary solution to the above problem is outlined in Montanari and Doretti (2023). However, the method described in that note depends on an arbitrary threshold parameter. This flaw

is the key fact motivating the novel approach proposed in this paper, which consists of two steps. First, a new probability distribution is derived which is specifically designed for Φ-transformed data from a Normal random variable with expectation equal to zero and standard deviation σ. In this way, overdispersion is explicitly addressed. Because of its similarity to the unit-interval Beta distribution, we name such a distribution "Betoidal", with characteristic parameter σ. In the second step, coarsening due to rounding is addressed by linking a multinomial distribution to the underlying Betoidal density, thereby obtaining what we term the "Beto-multinomial" model.

The rest of the paper is organized as follows. In Section 2, we introduce the Betoidal distribution and show how, in the absence of coarsening, maximum likelihood (ML) estimation for its characteristic parameter σ would match the well-established Normal framework. In Section 3, we provide details about how coarsening is handled and describe the ML estimation of the Beto-multinomial model. Such a model is then applied to the 2017 ANVUR data, for which results are presented in Section 4. In Section 5, some concluding remarks are given, and possible directions for adjusting the overall performance indicator ISPD as well as for future developments are sketched.

2. Betoidal distribution

Let Y be a Normal random variable with null expectation and standard deviation equal to σ, and consider the transformation $X = \Phi(Y)$. Then, X lies in the $(0\,;1)$ interval and, by standard results on densities of transformations, its probability density function (PDF) is given by

$$f_X(x;\sigma) = \frac{\sqrt{2\pi}}{\sigma}\phi\left(\frac{\Phi^{-1}(x)}{\sigma}\right)\exp\{[\mathrm{erf}^{-1}(2x-1)]^2\}, \tag{1}$$

where $\phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are, respectively, the PDF and the quantile function of the standard normal distribution and $\mathrm{erf}^{-1}(\cdot)$ is the inverse of the error function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt.$$

The CDF of X is

$$F_X(x;\sigma) = P(X \le x) = P(\Phi(Y) \le x) = P\big(Y \le \Phi^{-1}(x)\big) = F_Y\big(\Phi^{-1}(x)\big)$$

$$= F_Y\left(\sqrt{2}\,\mathrm{erf}^{-1}(2x-1)\right) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mathrm{erf}^{-1}(2x-1)}{\sigma}\right)\right], \tag{2}$$

whereas the quantile function is given by

$$F_X^{-1}(q;\sigma) = \frac{1 + \mathrm{erf}\{\sigma \cdot \mathrm{erf}^{-1}(2q-1)\}}{2}$$

for any $q \in (0,1)$. In the special case of $\sigma = 1$, the distribution of X is uniform, since for all $x \in (0\,;1)$ we have $f_X(x;1) = 1$ and $F_X(x;1) = x$, while for any $q \in (0,1)$, $F_X^{-1}(q;\sigma) = q$ holds. Conversely, $f_X(x;\sigma)$ is U-shaped for $\sigma > 1$ and bell-shaped for $\sigma < 1$, in a way very similar to the density of a unit-interval $\mathrm{Beta}(\alpha,\alpha)$ random variable, with $\alpha < 1$ and $\alpha > 1$, respectively. For these reasons, we refer to $f_X(x;\sigma)$ as the "Betoidal" distribution, i.e., $X \sim \mathrm{Betoidal}(\sigma)$.

Given a sample $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ obtained from n independent draws from a random variable $X \sim \mathrm{Betoidal}(\sigma)$, the log-likelihood function can be written as

$$\ell_X(\sigma; \boldsymbol{x}) = n \cdot \log\frac{\sqrt{2\pi}}{\sigma} + \sum_{i=1}^{n} \log\phi\left(\Phi^{-1}(x_i)/\sigma\right) + \sum_{i=1}^{n} [\mathrm{erf}^{-1}(2x_i - 1)]^2,$$

from which it is immediate to verify that

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} [\Phi^{-1}(x_i)]^2}$$

is the ML estimate of σ, since it is the unique solution of $\partial\ell_X(\sigma; \boldsymbol{x})/\,\partial\sigma = 0$, and it satisfies

$$\frac{\partial^2 \ell_X(\sigma; \boldsymbol{x})}{\partial^2 \sigma}\Big|_{\sigma=s} < 0.$$

Also, relying on standard ML theory, it is possible to compute the Betoidal expected Fisher information

$$\mathcal{I}_X(\sigma) \equiv -nE\left(\frac{\partial^2}{\partial^2\sigma}\log f_X(X;\sigma)\right) = \frac{2n}{\sigma^2}.$$

Denoting by S the random equivalent of s (that is, the ML estimator of σ), from the above expression it is immediate to derive the variance

$$V(S) = \mathcal{I}_X(\sigma)^{-1} = \sigma^2/(2n).$$

As expected, these results agree with the well-established theory of ML estimation for normal data. In particular, reverting to the normal scale leads to the usual sample standard deviation estimator $S = (\sum_i^n Y_i^2/n)^{-1/2}$, which is known to be the ML estimator of σ when the observed data $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ are realizations of n independent draws from a random variable $Y \sim N(0, \sigma^2)$.

3. Addressing coarsening

In the presence of a coarsening mechanism like the one taking place within the ANVUR procedure, a random sample $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ obtained from n independent draws from $X \sim \mathrm{Betoidal}(\sigma)$ is not directly available. In fact, given an underlying realization x and a width rounding interval $\lambda > 0$ such that $J = \lambda^{-1}$ is an integer number, the observed data is the vector $\boldsymbol{r} = (r_1, \dots, r_n)^\top$, where r_i is the λj value for which $x_i \in (\lambda(j - 0{,}5)\ ;\ \lambda(j + 0{,}5)]$ $(j = 0, 1, \dots, J)$. For the ANVUR data, the λj values are 0.000, 0.005, 0.010, ..., 0.990, 0.995, 1.000, and $J = 200$. Ultimately, $\boldsymbol{r}$ is a random sample obtained from *n* independent draws from a discrete random variable R with support points λj $(j = 0, 1, \dots, J)$ and probability distribution

$$p(\lambda j; \sigma) := P(R = \lambda j; \sigma) = \int_{\lambda(j-0.5)}^{\lambda(j+0.5)} f_X(x; \sigma)dx = F_X[\lambda(j + 0.5); \sigma] - F_X[\lambda(j - 0.5); \sigma].$$

As a consequence, denoting by n_j the observed frequency of the value λj, so that $\sum_{j=0}^{J} n_j = n$, we

have that $\boldsymbol{n} = \left(n_0, \ldots, n_J\right)^\top$ is a realization of a multinomial random variable $\boldsymbol{N} = \left(N_0, \ldots, N_J\right)^\top$ with n draws and success probability vector

$$\boldsymbol{p}(\sigma) := \left(p(\lambda 0; \sigma), p(\lambda 1; \sigma), \ldots, p(\lambda J; \sigma)\right)^\top.$$

The probability of such a realization is given by

$$f_N\left(\boldsymbol{n} \mid n, \boldsymbol{p}(\sigma)\right) = \frac{n!}{\prod_{j=0}^{J} n_j\,!} \prod_{j=0}^{J} p(\lambda j; \sigma)^{n_j}.$$

Since the elements of the success probability vector $\boldsymbol{p}(\sigma)$ are directly linked to the Betoidal PDF and CDF, we refer to the above model as to the Beto-multinomial model. To estimate σ from such a model, it is immediate to observe that the log-likelihood of the observed frequency vector $\boldsymbol{n}$ is given by

$$\ell_N(\sigma; \boldsymbol{n}) = k + \sum_{j=0}^{J} n_j \ln p(\lambda j; \sigma), \tag{3}$$

where k is a constant not depending on σ. In this setting, maximization of (3) can be performed via standard numerical procedures such as the Newton-Raphson (NR) algorithm. In detail, in order to obtain the ML estimate $\hat{\sigma}$, one can start from an initial guess $\sigma^{(0)}$ and iterate until convergence the update formula

$$\sigma^{(t+1)} = \sigma^{(t)} - \{h_N\left(\sigma^{(t)}; \boldsymbol{n}\right)\}^{-1} s_N\left(\sigma^{(t)}; \boldsymbol{n}\right),$$

where $s_N(\sigma; \boldsymbol{n})$ and $h_N(\sigma; \boldsymbol{n})$ are the score function and its derivative, respectively. The standard error of $\hat{\sigma}$ can be calculated by taking the square root of the inverse of the multinomial observed Fisher Information computed at convergence, i.e.,

$$\mathrm{se}_o(\hat{\sigma}) = \sqrt{-\frac{1}{h_N(\hat{\sigma}; \boldsymbol{n})}}.$$

Alternatively, one can rely on the multinomial expected Fisher information

$$\mathcal{I}_N(\sigma) = -n\mathrm{E}\left(\frac{\partial^2}{\partial^2 \sigma} \log f_N\left(\boldsymbol{N}; n, \boldsymbol{p}(\sigma)\right)\right)$$

to obtain $\mathrm{se}_e(\hat{\sigma}) = \mathcal{I}_N(\hat{\sigma})^{-1/2}$.

4. Results for the ANVUR 2017 data

As mentioned in Section 1, for the 2017 ANVUR evaluation exercise the distribution of the performance index ISPD across academic departments, r_i for $i = 1, \ldots, D = 766,$ is heavily U-shaped. In fact, 119 out of the 766 departments are marked with the worst performance level (that is, $r_i = 0$), and the same number are marked with the best one (i.e., $r_i = 1$). Such a symmetrical behavior extends to the next two performance levels: there are 29 departments with $r_i = 0.005$, the same number of those with $r_i = 0.995$. These frequencies are considerably higher than the remaining ones, as shown by the histogram in Figure 1a.

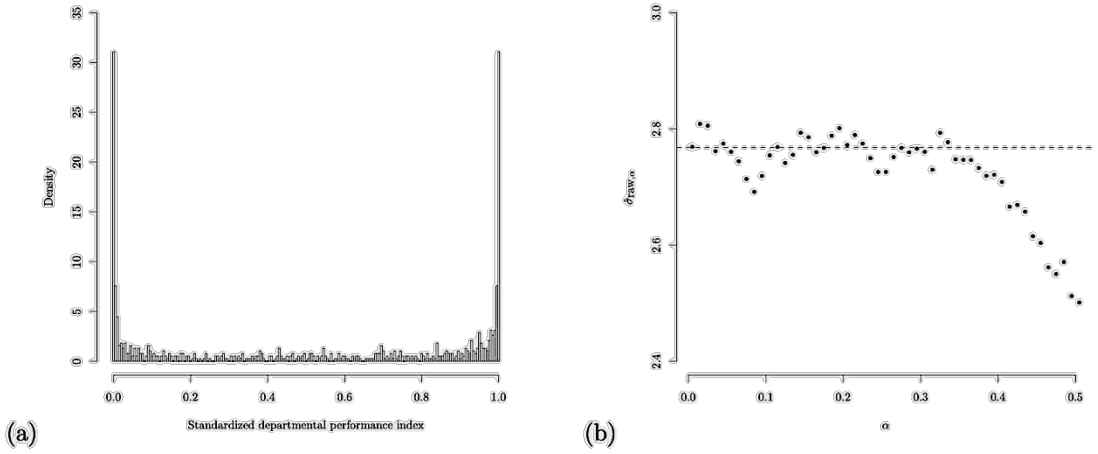


Figure 1: (a) Histogram of the standardized departmental performance index; 2017 ANUVUR data. (b) Heuristic approach to estimate σ as a function of the threshold $\alpha \in [0.005, 0.505]$ (with step 0.01). The dashed line corresponds to $\hat{\sigma} = 2.7677$.

The above pattern provides evidence to conclude that the standard deviation σ of the underlying probability distribution of non-transformed scores y is greater than 1. In this regard, it has to be noticed that - due to the presence of intra-departmental correlation and different sizes of departments - the probability distribution underlying the y scores is a *mixture* of D normal components with a zero mean and department-specific standard deviations, σ_i, rather than a normal variate itself. However, such a mixture distribution is well approximated by a normal distribution with a zero mean and variance $\sigma^2 = D^{-1}\sum_{i=1}^{D}\sigma_i^2$. For the ANVUR 2017 data, such an approximation works quite well (again, we refer for details to Montanari and Doretti, 2023). Then, within this setting, it is possible to use the Beto-multinomial model, as described in Section 3, to obtain the ML estimate of the parameter σ. From the ANVUR 20217 data the standard deviation of y is estimated to be $\hat{\sigma} = 2.7677$, with the two standard errors $\mathrm{se}_e(\hat{\sigma})$ and $\mathrm{se}_o(\hat{\sigma})$ being equal up to the fourth digit (0.0932). NR algorithm convergence is quite fast, with the likelihood maximization taking less than 15 iterations for a range of plausible starting values of σ (from 0.6 to 4; notice that numerical issues might occur with more extreme values). These results are essentially in line with the preliminary approach in Montanari and Doretti (2023), where for a chosen threshold $\alpha \in [0.005\ ;\ 0.505]$ the estimate of σ is given by

$$\hat{\sigma}_{\mathrm{raw},\alpha} = \frac{Z_{\alpha/2}}{\Phi^{-1}\left(1 - 0.5\hat{P}_\alpha\right)},$$

being $-Z_{\alpha/2}$ the standard normal $\alpha/2$ quantile and $\hat{P}_\alpha$ the sample proportion of departments with an ISPD lower than $\alpha/2$ or greater than $1 - \alpha/2$. Hence, the Beto-multinomial model appears to be more stable than this heuristic method, which depends on the arbitrary choice of α as shown in Figure 1b.

5. Conclusions

The ISPD values computed by the Italian governmental agency ANVUR to rank academic departments are highly concentrated in the tails, with almost one third of the departments assuming the extreme values 0 or 1 in the 2017 ranking (in the 2022 ranking this phenomenon is even more pronounced). This is because each department's ISPD is the rounded value of the standard normal CDF transformation applied to data from an approximately normal distribution with a mean equal

to zero and standard deviation $\sigma > 1$. In this work, we cope with the problem of estimating σ using the rounded ISPD values only. In this respect, we improve the heuristic approach in Montanari and Doretti (2023) via a more analytical method based overall ISPD frequency distribution, rather than on the simple proportion of departments whose values are more extreme than an arbitrary threshold.

Our solution involves a novel probability distribution that we name "Betoidal" because of its similarity to the unit-interval symmetric beta distribution. The characteristic parameter of this novel distribution is indeed the standard deviation of interest, σ. We have shown how ML theory for σ matches the well-known results holding for the normal distribution and, more importantly, how estimation can be performed even in a setting where coarsening occurs. This is the case for the application at hand, where coarsening is due to rounding. In particular, such an issue is addressed by linking a multinomial distribution to the underlying Betoidal density, thereby obtaining the "Beto-multinomial" model.

For the 2017 ANVUR data, this model leads to the estimate $\hat{\sigma} = 2.7677$, with a standard error equal to 0.0932. From this estimate, department-specific standard deviations can be derived under the assumption that the intra-departmental correlation is constant. Under the latter, such a correlation is given by $\hat{\rho} = (\hat{\sigma}^2 - 1)/(\bar{N} - 1) = 0.0514$, where $\bar{N}$ is the average number of evaluated papers across the D departments, and the estimated department specific standard deviation would be $\hat{\sigma}_i = [1 + \hat{\rho}(N_i - 1)]^{0.5}$, for $i = 1, \dots, 766$. Thus, due to the different sizes of the departments, N_i, these standard deviations range from 1.48 to 4.98. Indeed, this different scaling among departments is what compromises the overall fairness of the ISPD indicator, as stated in Section 1: if two departments are associated to distinct underlying standard deviations, the same ISPD value actually reflects two different performance levels (Montanari and Doretti, 2023). In this respect, a more appropriate measure for comparing departments would be $\Phi(y_i/\hat{\sigma}_i)$. But, de-coarsened $\Phi(y_i)$ values would be needed in order to revert to y_i and recalculate $\Phi(y_i/\hat{\sigma}_i)$. Besides, access to the standardized scores attributed to each scientific product would enable removing the assumption of constant intra-departmental correlation.

Future developments might involve relying on proper de-coarsening techniques to fully exploit information about the curvature of the Betoidal density within each rounding interval. Moreover, the proposed approach could be extended to settings where rounded performance indices lying in the [0 ; 1] interval are generated by an underlying parametric model different from the Normal distribution.

References

ANVUR (2022). Nota metodologica sul calcolo dell'indicatore ISPD. Technical report. *https://www.anvur.it/wp-content/uploads/2018/04/Nota_metodologica_ISPD_AN_.pdf*

Montanari, G.E., Doretti, M. (2023). On the impact of intraclass correlation in the ANVUR evaluation of academic departments. In *SIS 2023 Statistical Learning, Sustainability and Impact Evaluation.* Book of short papers, pp. 417–421. Pearson.

ROARS (2022). VQR: la lista segreta dei 120 dipartimenti con zero in pagella. Technical report. *https://www.roars.it/vqr-la-lista-segreta-dei-120-dipartimenti-con-zero-in-pagella/*

doi.org/10.26398/asaproc.00124

Design of a test for assessing teamwork and collaborative skills with competence-based knowledge space theory

Federica Morleo[ab], Alessandra Vitanza[b], Pasquale Anselmi[a].

a. Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Italy.
b. Institute of Cognitive Sciences and Technologies, National Research Council, Italy.

1. Introduction

Nowadays, social changes due to information technology development and globalization are becoming new challenges for education (OECD, 2020a). In response, modern individuals require flexible, transformative, and adaptive skills to thrive successfully in our world. Education needs to address global issues by fostering a spectrum of attitudes, values, and behaviors to provide sustainable competencies for all learners (UNESCO, 2017). This has become an urgent matter in the school system, hence increasing attention on the quality of teachers' professionalism, for being empowered to contribute to the ambitious aim of the UN Sustainable Development Goals (UN, 2015). A focus on the teaching quality is given by Teaching and Learning International Survey (TALIS; OECD, 2020b), and more recently enhanced by the OECD project (OECD, 2023). They identify five key elements of a teacher's professionalism: (a) the development of knowledge and skills base, (b) the range of quality training and lifelong learning, (c) the opportunity to promote collaboration, (d) the power of being responsible and autonomous in policy and practice, and (e) the prestigious status of the profession. Throughout their map, different possible ways of quality teacher improvement can be exploited. Taking into consideration the focus on teachers' professional collaboration given by the last TALIS (OECD, 2020b), we decided to focus our attention on collaboration and teamwork to improve teachers' professionalism. The idea of teacher collaboration and teamwork is strongly related to school improvement and effectiveness; indeed, it is considered an essential pillar of teachers' professionalism. Collaboration requires skills to single out solutions to complex challenges and to deal with networks, while managing the work across organizational boundaries (OECD, 2021). Moreover, in today's educational context, the trend is moving towards cooperative learning to meet the needs of the workforce in society, hence these kinds of skills among teachers are even more relevant, as they can affect collaborative skills among students (OECD, 2020c): another important goal for the future of education.

The first step toward effectively improving skills is assessment, which provides precise feedback about the strengths and weaknesses of individuals. This is the core value of formative assessment, which allows to modulate interventions for obtaining excellence for all. However, a critical issue within the educational system is the availability of assessment tools capable of accurately identifying the skills mastered by an individual and those that are not. In this framework, competence-based knowledge space heory (CbKST; Doignon, 1994; Heller et al., 2013) is a mathematical theory for the individual assessment of skills that provides a precise description of the set of skills an individual possesses within a specific domain, based on the responses given to the items of a test. A test could be designed from scratch within CbKST, by constructing items that require specific skills to be solved. To date, applications of CbKST to test development are less common than those of Classical Test Theory and Item Response Theory, and the existing ones mainly focus on hard skills (Anselmi et al., 2012, 2017; de Chiusole et al., 2020; Robusto et al., 2010).

This work deals with the construction and validation of a test for the assessment of collaborative and teamwork skills in pre-primary and primary school teachers in the framework of CbKST.

2. Method

Development of the test

Exploring collaborative and teamwork skills for teachers is a hard challenge because teaching is a multifaceted profession. Despite the difficulty in describing the complexity of teamwork skills, two domains can be found: the former related to the relationship area and the latter to the processes around problem-solving. The interpersonal domain comprises empathy, communication, and getting along well with others (Zhou et al., 2023); whereas the problem-solving domain includes, among others, skills like decision-making, process control, and adjustment, a sense of responsibility, and perseverance (Awuor et al., 2022). Accordingly, the present work categorizes teamwork skills into two dimensions, named respectively "relationship domain" and "work management domain". Specifically, considering the goal of this study — to create a multidimensional assessment tool for individuals' skills based on CbKST — the choice of skills is driven by the need to focus on those that are observable in individuals' behaviors, resulting from an objective assessment. In this sense, the difficulty in defining teamwork and collaborative skills is reflected in the limited international research on learning and assessing life skills (Care et al., 2018). Within the relationship domain, this work focuses on "effective communication" and "conflict resolution", whereas, in the work management domain, the attention is on "action planning "and "performance monitoring and feedback", as these skills are distinctive aspects of teachers' professionalism. Moreover, TALIS explores a broad vision of teamwork activities among teachers (OECD, 2020b), underlying that they have many opportunities to interact with others, with different levels of interdependence and coordination (OECD, 2020c). Table 1 provides the skills chosen for the study and a detailed description and examples of each. The skills are assumed to be independent.

Table 1: Description of teamwork and collaborative skills in teachers

Domains	Skills	Description
Relationship domain	Effective Communication	Using proactive communication strategies (e.g., proposing weekly task updates). Using active listening techniques (e.g., paraphrasing a colleague's idea). Managing verbal and non-verbal instructions (e.g., clarifying instructions with gestures).
	Conflict Resolution	Recognizing type and relevance of a conflict situation (e.g., identifying task role confusion). Identifying conflict resolution strategies (e.g., suggesting collaborative problem-solving). Using negotiating approaches, dealing with others (e.g., balancing workload with colleagues).
Work Management domain	Action Planning	Identifying a shared objective (e.g., focusing on completing a project). Identifying and planning action strategies (e.g., creating group task timeline). Identifying and establishing roles and tasks equally (e.g., assigning roles based on skills).
	Performance Monitoring and Feedback	Recognizing the obstacles and errors during group work (e.g., noticing unclear instructions). Providing individual and collective feedback (e.g., highlighting students progresses). Identifying problem-solving strategies (e.g., suggesting specific pedagogical strategies).

The test we created consists of 20 items, formulated both in multiple-choice and short free format. Items are questions that teachers are asked to answer, by making decisions with colleagues and participating in various school activities. The test can be found at: https://osf.io/57e6w/.

In the framework of CbKST, each item needs to be associated with the skills that are assumed to be necessary for solving it. Table 2 shows the item-skill assignment of the test, in which each item is associated only with one skill. For instance, Item 1 is assumed to require skill "action planning" to be solved; Item 18 is assumed to require skill "conflict resolution". There are 5 items for each of the four skills.

Table 2: Preliminary item-skill assignment

	Skills			
Item	Effective Communication	Conflict Resolution	Action Planning	Performance Monitoring and Feedback
1			X	
2	X			
3			X	
4			X	
5		X		
6			X	
7	X			
8		X		
9	X			
10				X
11	X			
12	X			
13		X		
14		X		
15				X
16				X
17			X	
18		X		
19				X
20				X

Participants

The empirical validation of the item-skill assignment to real data involved a sample of 139 school teachers from pre-primary and primary schools in Italy, who compiled the test after the approval of their school headteacher. Teachers individually filled out the test with their computer or smartphone.

Analytic approach

Within CbKST, the item-skill assignment needs to be validated empirically. The probabilistic model that we have used to this aim is a competence-based extension of the basic local independence model (BLIM; Falmagne & Doignon, 1988a, 1988b) denoted as competence-based local independent model (CBLIM; Heller et al., 2015). The validation involved a complex and iterative process, based on the results of the application of the CBLIM (a detailed description of the computational procedures can be found in Anselmi et al., 2022). First, the absolute goodness-of-fit based on the item-skill assignment was evaluated. In this work, the absolute goodness-of-fit was considered satisfactory if p-value was > .10. Second, the local identifiability of the model parameters was ascertained. Third, in the case of satisfactory goodness-of-fit, the estimates of the careless error and lucky guess probabilities characterizing each item were interpreted. The first is the probability that an individual

fails the item despite mastering all the skills needed for solving the item. The second is the probability that an individual solves the item despite missing some or all the skills required by it. Possible reasons might be related to the properties of the items, resulting in inattention for the careless error, and guessing or cheating for the lucky guess. The estimates of careless error and lucky guess probabilities are particularly informative about the item-skill assignment. In this work, the careless error and lucky guess probabilities of an item were considered satisfactory if their value was < .50 and their sum was < .10. A high careless error for an item may indicate that the item requires more skills than expected or that it has been associated with one or more superfluous solution strategies. A high lucky guess may indicate that the item requires an alternative solution strategy or that it has been associated with more skills than needed. Starting from the item-skill assignment in Table 1, created based on the skills assumed to underlie each item, various progressive adjustments were developed considering the outcomes listed above.

3. Results

The CBLIM based on the item-skill assignment in Table 2 is locally identifiable but the *p*-value is statistically significant (*p*-value .07), thus suggesting that the item-skill assignment does not describe data satisfactorily. Consequently, an alternative item-skill assignment was constructed based on the outcomes of goodness-of-fit and the estimates of careless error and lucky guess parameters. It is depicted in Table 3, in which the main change concerns a further categorization applied to the skills "effective communication" and "conflict resolution". Specifically, the skill "effective communication" was decomposed in the two independent skills: "using proactive communication strategies" and "using active listening techniques". The skill "conflict resolution", instead, was decomposed in the two skills: "recognizing type and relevance of a conflict situation", and "identifying conflict resolution strategies". In addition, some items previously associated with only one skill are now associated with two skills, in two different ways. When the item is associated with two skills in the same row, it means that the item needs both skills to be solved, thus there is only one solution strategy, involving the mastering of both skills. When the item is associated with two skills in two different rows, it means that the item needs either one or the other skill to be solved, thus there are two solution strategies, each involving one of the two skills. For instance, Item 4 needs either skill "identifying and planning action strategies" or skill "monitoring performance and providing feedback" to be solved; whereas Item 5 needs both skills "recognizing type and relevance of a conflict situation" and "identifying conflict resolution strategies" to be solved.

The CBLIM based on the new item-skill assignment is locally identifiable and shows a satisfactory fit to the data (*p*-value .37).

Table 3 also reports the estimates of the careless error and lucky guess probabilities for each item. In general, the mean of careless error probabilities (.21) is less than that of the lucky guess probabilities (.37). Specifically, the careless error probability is high for Items 12 (.74) and 17 (.51). The lucky guess is rather high for Items 2 (.54), 4 (.71), 9 (.52), 16 (.68), 19 (.61). On the whole, according to Section "Analytic approach", careless error and lucky guess probabilities could indicate that the item-skill assignment needs further refinement. However, considering the small sample size, the results suggest that the new item-skill assignment is effectively more appropriate than the preliminary assignment.

Table 3 New item-skill assignment with the estimates of items careless error (CE) and lucky guess (LG) probabilities

Item	Skills						CE	LG
	Using proactive communication strategies	Using active listening techniques	Recognizing type and relevance of conflict situation	Identifying conflict resolution strategies	Identifying and planning action strategies	Monitoring performance and providing feedback		
1					X		.23	.40
2			X				.00	.54
3					X		.47	.18
4						X	.07	.71
4					X			
5			X	X			.26	.31
6						X	.05	.45
6					X			
7	X						.00	.13
8			X	X			.43	.48
9		X					.04	.52
10						X	.23	.05
11	X						.29	.00
11		X						
12	X	X					.74	.12
13				X			.20	.43
13			X					
14			X				.05	.35
14				X				
15						X	.06	.13
16					X		.10	.68
16						X		
17					X	X	.51	.37
18				X			.21	.36
19					X		.32	.61
19						X		
20						X	.00	.48

4. Discussion and conclusion

The present work described the development from scratch of a test for assessing soft skills in pre-primary and primary school teachers in the framework of CbKST. Results concerning the interpretation of careless error and lucky guess probabilities of the items led us to modify the preliminary item-skill assignment, including new skills that were not previously considered, and associating some items with different solution strategies. A core aspect of the study was to explore how to interpret the estimates of item careless error and lucky guess probabilities to guide appropriate modifications for improving test validity.

However, it is essential to acknowledge a limitation of the study, which involves a sample of only 139 teachers, making it challenging to generalize the results. Consequently, data collection is still in progress.

References

Anselmi, P., Robusto, E., Stefanutti, L. (2012). Uncovering the best skill multimap by constraining the error probabilities of the gain-loss model. *Psychometrika*, **77**, pp. 763-781. https://doi.org/10.1007/s11336-012-9286-0.

Anselmi, P., Stefanutti, L., de Chiusole, D., Robusto, E. (2017). The assessment of knowledge and learning in competence spaces: The gain–loss model for dependent skills. *British Journal of Mathematical and Statistical Psychology*, **70**(3), pp. 457-479. https://doi.org/10.1111/bmsp.12095.

Anselmi, P., Heller, J., Stefanutti, L., Robusto, E. (2022). Constructing, improving and shortening tests for skill assessment. *Journal of Mathematical Psychology*, 106, 102621. https://doi.org/10.1016/j.jmp.2021.102621.

Awuor, N.-O., Weng, C., Piedad, E. J., Militar, R. (2022). Teamwork competency and satisfaction in online group project-based engineering course: The cross-level moderating effect of collective efficacy and flipped instruction. *Computers & Education*, **176**, 104357. https://doi.org/10.1016/j.compedu.2021.104357.

Care, E., Kim, H., Vista, A., Anderson, K. (2018). *Education System Alignment for 21st Century Skills: Focus on Assessment*. Washington, DC. Brookings Institute.

de Chiusole, D., Stefanutti, L., Anselmi, P., Robusto, E. (2020). Stat-Knowlab. Assessment and learning of statistics with competence-based knowledge space theory. *International Journal of Artificial Intelligence in Education*, **30**, pp. 668-700. https://doi.org/10.1007/s40593-020-00223-1.

Doignon, J.-P. (1994). Knowledge spaces and skill assignments. In G. H. Fisher, D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics and Methodology* (pp. 111-121). Springer-Verlag.

Falmagne, J.-C., Doignon, J.-P. (1988a). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, **41**(1), pp. 1–23. https://doi.org/10.1111/j.2044-8317.1988.tb00884.x.

Falmagne, J.-C., Doignon, J.-P. (1988b). A Markovian procedure for assessing the state of a system. *Journal of Mathematical Psychology*, **32**, pp. 232–258. https://doi.org/10.1016/0022-2496(88)90011-9.

Heller, J., Stefanutti, L., Anselmi, P., Robusto, E. (2015). On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika*, **80**(4), pp. 995–1019. https://doi.org/10.1007/s11336-015-9457-x.

Heller, J., Ünlü, A., Albert, D. (2013). Skills, competencies and knowledge structures. In J.-C. Falmagne, D. Albert, C. Doble, D. Eppstein, X. Hu (Eds.), *Knowledge Spaces: Applications in Education* (pp. 229-242). Springer-Verlag.

Organization for Economic Co-operation and Development (2023). Teaching for the future: Global engagement, sustainability and digital skills, international summit on the teaching profession. *OECD Publishing*. https://doi.org/10.1787/d6b3d234-en.

Organization for Economic Co-operation and Development (2021). Teaching as a knowledge profession: studying pedagogical knowledge across education systems, educational research and innovation. *OECD Publishing*. https://doi.org/10.1787/e823ef6e-en.

Organization for Economic Co-operation and Development (2020a). OECD future of education and skills 2030: OECD Learning Compass 2030. *OECD Publishing*.

Organization for Economic Co-operation and Development (2020b): TALIS 2018 Results (Volume II). In Teachers and school leaders as valued professionals, TALIS. *OECD Publishing*. https://doi.org/10.1787/19cf08df-en.

Organization for Economic Co-operation and Development (2020c). Teaching in Focus. Professional collaboration for teachers working in challenging environments. *OECD Publishing*. https://doi.org/10.1787/c699389b-en.

Robusto, E., Stefanutti, L., Anselmi, P. (2010). The gain-loss model: A probabilistic skill multimap model for assessing learning processes. *Journal of Educational Measurement*, **47**(3), pp. 373-394. https://doi.org/10.1111/j.1745-3984.2010.00119.x.

United Nations (2015). *Sustainable Development Goals*. General Assembly Resolution, A/RES/70/1. United Nations.

United Nations Educational, Scientific and Cultural Organization (2017). *Education for Sustainable Development Goals. Learning Objectives*. Unesdoc Digital Library.

Zhou, W., Dong, J., Zhang, W. (2023). The impact of interpersonal interaction factors on consumers' purchase intention in social commerce: a relationship quality perspective. *Industrial Management & Data Systems*, **123**(3), pp. 697-721. https://doi.org/10.1108/IMDS-06-2022-0392.

doi.org/10.26398/asaproc.00125

Efficiency of Italian courts: Awareness from spatial stochastic frontier analysis

Eugenia Nissi[a], Annalina Sarra[b], Adelia Evangelista[b], Tonio Di Battista[b]
[a] Department of Economics, "G. d'Annunzio" University of Chieti-Pescara, Italy
[b] Department of Socio-Economic, Managerial, and Statistical Studies, "G. d'Annunzio" University of Chieti-Pescara, Italy

1. Introduction

The efficiency of judicial systems is crucial for effective governance and economic stability. As institutions responsible for administering justice, courts play an essential role not only in upholding legal standards but also in impacting broader socio-economic outcomes. Recent years have seen substantial efforts to analyze and enhance the performance of judicial systems worldwide, with a particular focus on optimizing their efficiency. In Italy, the judicial system has consistently been identified as one of the least efficient in Europe, according to various international assessments (see, for instance CEPEJ, 2016).

Persistent case backlogs and lengthy trials contribute to delays in justice, often exacerbated by complex procedures and bureaucratic inefficiencies. Limited resources, including a shortage of judges and courtrooms, further hinder case resolution. In addition, the complexity of the legal system, with its multiple levels and jurisdictions, adds to the prolonged nature of legal proceedings. This perceived inefficiency raises concerns about its impact on economic development, financial market stability, and general public trust in legal institutions. Consequently, there is a growing need to address these issues, which has spurred research into more accurate and insightful methods for evaluating judicial performance.

Various approaches have been developed to assess judicial efficiency, each focusing on different aspects of court performance. Common metrics include the time required to resolve a case, the number of cases adjudicated by a court, technical efficiency scores, and case clearance rates (Ippoliti, 2015). The methods used to benchmark judicial performance range from basic to complex. Econometric models, such as those applied by Beenstock and Haitovsky (2004), offer foundational insights into efficiency levels. Other techniques include the free disposal hull (Tulkens, 1993), which provides a non-parametric approach to measuring efficiency, and the directional distance function (Falavigna et al., 2015), which evaluates efficiency relative to a defined frontier. Data envelopment analysis (Schneider, 2005) and Malmquist indices (Falavigna et al., 2018) are also commonly employed, offering a thorough assessment of productivity changes over time. Moreover, integrating spatial and temporal dimensions into efficiency models can deepen the understanding of how external factors, such as regional disparities or variations in case types, influence judicial efficiency.

This study adopts a spatial stochastic frontier model to examine the efficiency of Italian courts. The model integrates spatial dependencies, allowing us to account for regional variations and other external factors that can influence court performance. By incorporating spatial elements, our approach aims to provide a more nuanced understanding of judicial efficiency across different regions and case types.

The remainder of this paper is structured as follows. First, we present the data used for the analysis, detailing the sources and characteristics of the judicial performance metrics. Next, we describe the methodology of the spatial stochastic frontier model and its application to our dataset. Then we discuss the results of the analysis, highlighting key findings and insights into the efficiency of Italian courts. Finally, we offer concluding remarks and policy recommendations based on our findings, discussing implications for judicial reform and potential areas for future research.

2. Data

To explore the operations of Italian courts, we analyzed a dataset provided by the Italian Ministry of Justice. This dataset, which refers to the year 2021, encompasses detailed information on pending cases, incoming cases, resolved cases, and the number of judges and auxiliary staff across 138 ordinary courts in Italy. Our analysis focuses on five factors: the number of judges (JJ), administrative staff (AA), incoming cases (II), pending cases (PP) and the average duration of cases (DD), measured in days. These factors are used to evaluate the output, specifically the number of resolved cases (SS).

Table 1 provides an overview of the descriptive statistics for Italian courts across different regions in 2021. The data reveals notable regional disparities in court operations, highlighting differences in the number of resolved cases, incoming cases, and pending cases. In particular, the North-West region stands out for resolving the highest number of cases, though with significant variability, indicating uneven performance. The North-East shows more consistency in its operations, managing pending cases more effectively, and having the shortest and most consistent case durations. In contrast, the Center and South regions exhibit substantial variability, especially in pending cases and case duration, suggesting possible backlog issues and inefficiencies. The Islands region, while resolving fewer cases overall, shows a steadier performance, though it also experiences longer case durations similar to the South.

The disparities in the number of judges and administrative staff across regions further underscore the uneven distribution of resources, with the North-West and Center showing the greatest variability. The North-East and Islands appear to have more balanced administrative support, potentially contributing to their more stable performance.

Globally, these findings suggest that resource allocation, caseload management, and regional characteristics significantly impact the efficiency of court operations in Italy. This variability calls for region-specific interventions to improve the overall efficiency and effectiveness of the judicial system.

Table 1: Descriptive statistics for Italian courts by macroregion. Year 2021.

Macroregion	Resolved Cases (SS)	Incoming Cases (II)	Pending Cases (PP)	Number of Judges (JJ)	Administrative Employees (AA)	Average Duration (DD) [days]
North-West	3077	2859	2441	42	129	1023
	(568 - 15067)	(488 - 13346)	(313 - 12911)	(8 - 290)	(34 - 730)	(704 - 1633)
North-East	2193	2093	1743	29	100	1038
	(447 - 4593)	(418 - 4252)	(228 - 3929)	(9 - 72)	(33 - 243)	(553 - 1362)
Center	2736	2596	2903	36	125	1336
	(533 - 21291)	(454 - 22657)	(775 - 22501)	(6 - 379)	(21 - 1199)	(682 - 1988)
South	2813	2245	2728	42	133	1658
	(422 - 17191)	(337 - 13989)	(424 - 17775)	(6 - 399)	(23 - 1032)	(870 - 2751)
Islands	1678	1486	2067	32	110	1705
	(129 - 7189)	(133 - 5989)	(227 - 8936)	(6 - 125)	(26 - 419)	(870 - 2845)

3. The stochastic frontier model with spatial dependence

The stochastic frontier analysis (SFA) model has emerged as a central tool to measure efficiency. SFA facilitates the estimation of parameters that define the frontier of the production function and offers a means to test the validity of the model.

The model, first introduced by Aigner et al. (1977) and Meeusen and van den Broeck (1977), is formulated as follows:

$$y = X\beta + vu, \tag{1}$$

where y denotes a vector of size $n \times 1$, containing log-transformed output observations for each unit in the sample (where $i = 1, \ldots, n$); the matrix X is $n \times k$ (including a constant term) and includes observations on $k - 1$ inputs. The parameter vector is represented by β, u denotes the inefficiency term (where $u \geq 0$), and v is the noise term.

The production technology is presumed to be log-linear, such as the Cobb-Douglas or translog functions. Generally, v is assumed to be identically and independently distributed (i.i.d.) across all units, often following a normal distribution. Similarly, the inefficiency term u, which is nonnegative, is typically assumed to follow either a half-normal or exponential distribution.

SFA presumes that the noise and inefficiency terms among units are independent (Coelli et al., 2005). However, this assumption may not hold in practice. Correlations among outcomes of different units can arise from endogenous effects (influence of peers), contextual effects (variation based on peers' characteristics), and correlated effects (similar behaviors due to unobserved characteristics or shared shocks) (see Manski, 2000). For example, in judicial systems, correlations may result from shared legal precedents, uniform procedural practices, or similar socio-economic contexts among courts or jurisdictions. Failing to account for these dependencies can lead to biased and unreliable efficiency estimates.

To address efficiency and productivity analysis with spatial dependencies, spatial stochastic frontier models (SSFMs) offer a sophisticated approach. SSFMs recognize that entities located near each other often exhibit similar efficiency or productivity levels due to common resources, policies, or regional economic conditions. By integrating elements of spatial econometrics with stochastic frontier models, SSFMs incorporate spatial lag or spatial error terms into their framework.

This study employs the model proposed by (Fusco and Vidoli, 2013), which integrates the spatial lag of inefficiency into the traditional stochastic frontier model. The relationship between output and input variables is established using the Cobb-Douglas production function. The model can be expressed in matrix form as follows:

$$\dot{y} = \dot{\boldsymbol{X}}\beta + \boldsymbol{v} - \boldsymbol{u} = \dot{\boldsymbol{X}}\beta + \boldsymbol{v} - (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\tilde{u} \tag{2}$$

where $\dot{y}$ is an $(n \times 1)$ vector of output in logarithmic form, $\dot{\boldsymbol{X}}$ is an $(n \times p)$matrix of inputs in logarithmic form, and $\beta$is a $(p \times 1)$ vector of parameters. The term v represents an $(n \times 1)$ vector of noise terms, where $v \sim iid\ N(0, \sigma_v^2 I)$, while u denotes an $(n \times 1)$vector of inefficiency terms with $u \sim N^{+}(0, (I - \rho W)^{-1}(I - \rho W^{T})^{-1}\sigma_u^2 I)$. The vector $\tilde{\boldsymbol{u}}$ is an $(n \times 1)$ random vector, where $\tilde{\boldsymbol{u}} \sim N(0, \sigma_{\tilde{\boldsymbol{u}}}^2 I)$, ρ is a spatial autoregressive coefficient with $|\rho| < 1$, and W is an $(n \times n)$ spatial weights matrix.

4. Main results

To evaluate the efficiency of courts, we start by applying a spatial stochastic frontier model, with the functional form of the frontier represented by the following equation:

$$\log(SS) = \beta_1 \log(II) + \beta_2 \log(PP) + \beta_3 \log(JJ) + \beta_4 \log(AA) + \beta_5 \log(DD) + v_i - (1 - \rho \sum w_i)^{-1} \tilde{u}_i\ . \tag{3}$$

Upon examining the estimates from the basic SFA model, summarized in Table 2, we observe that the standard deviation of inefficiency (σ_v^2 =0.0024) is statistically significant. The coefficients β_1 and β_5, along with the intercept, are statistically significant and exhibit signs that align with expectations. The coefficient β_1 is associated with the number of incoming cases. The positive sign of β_1 indicates that an increase in the number of incoming cases is associated with a corresponding increase in the number of resolved cases, assuming all other variables

remain unchanged. This reflects a direct relationship between the volume of incoming cases and the resolution output.

Conversely, the positive sign of β_5 signifies that an increase in the average duration of cases is associated with a higher number of cases resolved (SS). This suggests that longer case durations are linked to a greater volume of cases being resolved, potentially due to extended processing times or the allocation of additional resources to manage more cases effectively.

However, the traditional stochastic frontier model reveals spatial autocorrelation. Moran's I statistic, which is significant at 0.1390, indicates that spatial methodologies are appropriate for this dataset. The presence of significant spatial correlation in the SFA results justifies the use of the SSFA model to better account for spatial effects.

Table 3 presents the SSFA results using distance matrices with various cut-off distances. These results are robust across the different spatial weight matrices' specifications and consistent with those from the SFA model.

The efficiency levels calculated by the SSFA model, using a cut-off distance of 100 Km, are illustrated in Figure 1a. It is evident that courts in Rome and Sicily are less efficient than those in other Italian regions. This may stem from various factors, including a heavier workload, uneven resource distribution, and unique systemic inefficiencies. For example, Rome, being the capital, likely deals with more complex cases due to its political role. In Sicily, socio-economic issues and organized crime also contribute to a challenging workload. Imbalances in resource allocation are critical; insufficient judges and staff relative to case volume can lead to delays. Moreover, procedural complexities and outdated infrastructure may hinder operations.

Table 2: Estimation results by method

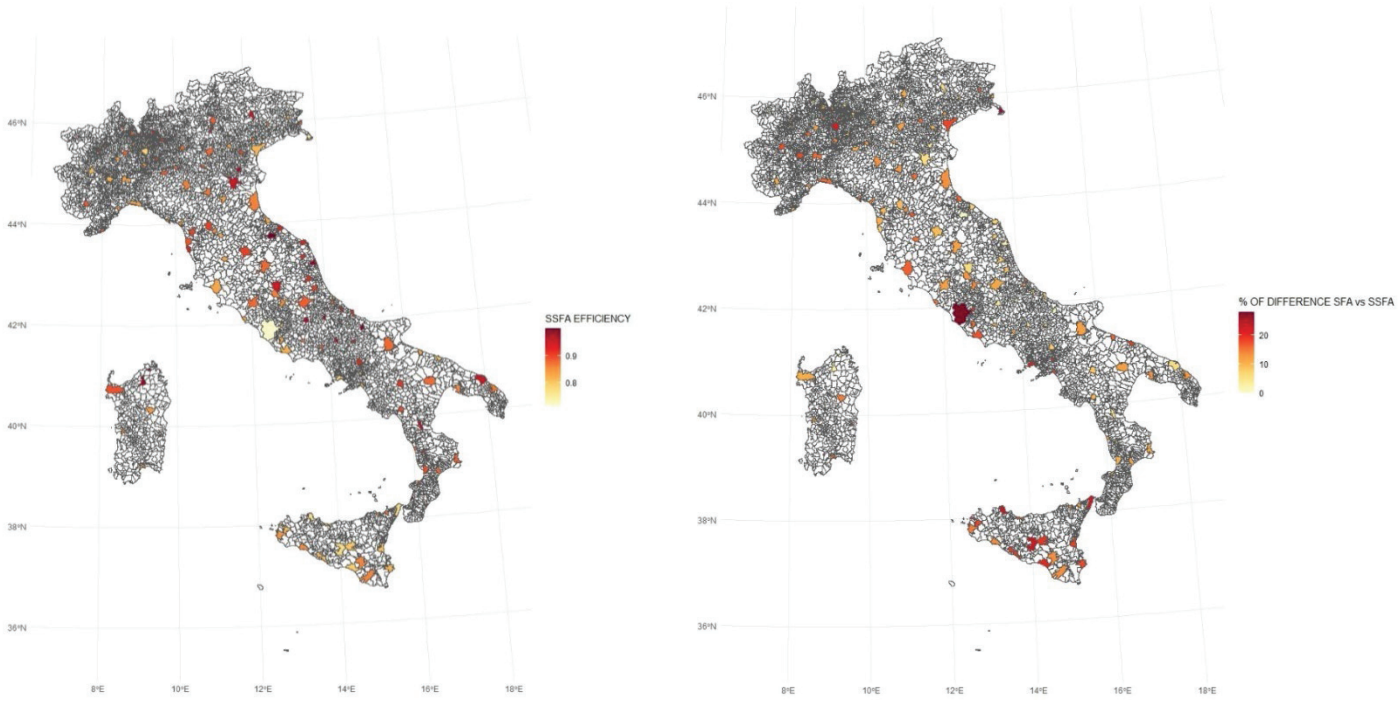
Variable	SFA	SSFA
Intercept	-0.0535***	-0.4691
β_1 - II	1.0059***	0.9252***
β_2 - PP	0.0032	0.0402
β_3 - JJ	0.0202	0.1621
β_4 - AA	-0.0219	-0.0192
β_5 - DD	0.1828***	0.1712**
σ_u	0.0000	0.0247***
σ_v	0.0024***	0.0000
Moran's I	0.0872*	-0.0682
ρ	-	-0.3742
Observations	138	138

Note: $***p < 0.001$; $** p < 0.01$; $* p < 0.1$; $p < 0.1$

To visually display local perturbations, we examine the differences in efficiency estimates between the SFA models with and without spatial interactions (i.e., Eff_{SFA}, Eff_{SSFA}). The efficiency index difference (x^*) is calculated as follows:

$$x^* = \frac{Eff_{SFA_i} Eff_{SSFA_i}}{Eff_{SSFA_i}} \times 100, \quad \forall\, i = 1, \ldots, n. \tag{4}$$

The results of this calculation are displayed in Figure 1b. Typically, the distance of the efficiency index (x^*) is positive because, in the SFA model, the spatial effect is incorrectly included in the error term, resulting in higher values compared to those estimated using the SSFA. Higher index values are consistently observed for the Rome court and for several courts in Sicily.



(a) (b)

Figure 1: (a) Efficiency of SSFA. (b) Percentage of difference between non-spatial versus spatial SFA.

Table 3: Estimation results for SSFA model with different distance matrices.

Variable	W110	W120	W130
Intercept	-0.4193.	-0.5831	-0.4271∗
β_1 - II	0.8989∗∗∗	0.9662∗∗∗	-0.8715∗∗∗
β_2 - PP	0.0475	0.0306	0.0592
β_3 - JJ	0.1806	0.1163	0.1018
β_4 - AA	-0.0200	-0.0162	0.0874
β_5 - DD	0.1650∗∗	0.1994∗∗	0.1489∗∗∗
σ_u	0.0223∗∗∗	0.0279∗∗∗	0.0200∗∗∗
σ_v	0.0000	0.0000	0.0000
Moran's I	-0.0649	-0.0489	-0.0477
ρ	-0.4377	-0.2538	-0.5373
Observations	138	138	138

Note: ∗∗∗p < 0.001, ∗∗p < 0.01, ∗p < 0.05, .p < 0.1.

5. Concluding remarks

This study analyzed the efficiency of Italian courts using a SSFA model, which considers regional differences and other external factors that influence judicial performance. The results suggest that traditional methods, such as SFA, do not adequately capture significant spatial dependencies that affect court efficiency. In particular, the analysis revealed that regional differences in court efficiency are strongly influenced by spatial factors. The courts in Rome and Sicily exhibit lower efficiency levels, attributable to a combination of regional structural

characteristics and resource distribution. The presence of heavy workloads, inadequate allocation of judges and staff, as well as procedural complexities, contribute to these inefficiencies. These findings highlight the importance of considering territorial specificities when planning interventions aimed at improving judicial performance. It is essential to further investigate spatial dependencies and procedural complexities to gain a better understanding of the underlying causes of the observed inefficiencies.

In both models, the parameter associated with the number of incoming cases, showed a positive and significant impact on court efficiency, indicating that an increase in workload correlates with higher productivity, likely due to better resource management. In contrast, the parameter linked to the average duration of the case, suggests that longer resolution times can lead to a higher number of resolved cases, reflecting the allocation of additional resources to handle backlogs.

In conclusion, this study provides valuable insights for policymakers seeking to enhance the effectiveness and efficiency of the Italian judicial system. By addressing regional disparities and focusing on the key drivers of inefficiency, targeted reforms can be implemented to improve the functioning of courts throughout the country.

References

Aigner, D., Lovell, C., Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, **6**, pp. 21–37.

Beenstock, M., Haitovsky, Y. (2004). Does the appointment of judges increase the output of the judiciary. *International Review of Law and Economics*, **24**, pp. 351–369.

Coelli, T.J., Prasada Rao, D.S., O'Donnell, C.J., Battese, G.E. (2005). *An Introduction to Efficiency and Productivity Analysis*. Springer, New York, 2005.

Council of Europe, European Commission for the Efficiency of Justice [CEPEJ] (2016). *European Judicial Systems: Efficiency and Quality of Justice* - CEPEJ Studies No. 23. Council of Europe Publishing, Strasbourg.

Falavigna, G., Ippoliti, R., Manello, A., Ramello, G.B. (2015). Judicial productivity, delay and efficiency: A directional distance function (DDF) approach. *European Journal of Operational Research,* **240**(2), pp. 592–601.

Falavigna, G., Ippoliti, R., Ramello, G.B. (2018). DEA-based Malmquist productivity indexes for understanding courts reform. *European Journal of Operational Research*, **62**, pp. 31–43.

Fusco, E., Vidoli, F. (2013). Spatial stochastic frontier models: Controlling spatial global and local heterogeneity. *International Review of Applied Economics*, **27**(5), pp. 679–694.

Ippoliti, R. (2015). La riforma della geografia giudiziaria: Efficienza tecnica e domanda di giustizia. *Economia Pubblica - Italian Journal of Public Economics*, **2**, pp. 91–124.

Manski, C. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, **14**(3), pp. 115–136.

Meeusen, W., van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, **18**, pp. 435–444.

Schneider, M. R (2005). Judicial career incentives and court performance: An empirical study of the German labour courts of appeal. *European Journal of Law and Economics*, **20**(2), pp. 127–144.

Tulkens, H. (1993). On FDH analysis: Some methodological issues and applications to retail banking, courts and urban transit. *Journal of Productivity Analysis*, **4**(1), pp. 183–210.

doi.org/10.26398/asaproc.00126

Campanian healthcare facilities under the microscope: Reviews, polarity, and georeferencing.

Simone Paesano[1], Maria GabriellaGrassia[1], Marina Marino[1], Rocco Mazza[3], Dario Sacco[1], Agostino Stavolo [1,2]

[1] Department of Social Sciences, University of Naples Federico II, Naples, Italy

[2]Faculty of Education, Free University of Bozen, Bozen, Italy

[3] Department of Political Sciences, University of Bari Aldo Moro, Bari, Italy

1. Introduction

The healthcare industry has become increasingly connected with online media, as patients routinely rate doctors and hospitals on social media platforms (Greaves et al., 2013). Despite the significance of using this data to improve patient care, many medical institutions have not yet capitalized on the online feedback available. This indecision may be partly due to uncertainty about how to effectively parse and analyse the large and unstructured dataset of feedback (Hariri et al., 2019; Liao et al., 2024). The present work aims to measure customer satisfaction and its distribution among the main hospitals in the Campania region. Using large language models (LLMs) as a semantic basis, we extracted 8.024 reviews to understand patients' perceptions and experiences with hospital services. We identify various healthcare facilities' prevailing sentiments, common criticisms, and strengths. LLMs will classify the reviews by polarity, categorizing them into positive, negative, and neutral groups. We propose developing a georeferenced visualization to map the distribution of positive and negative reviews for hospital companies across the Campania region. Finally, a comparison of three different models of sentiment analysis is proposed, with a focus on classification error.

2. Literature review

Consumers are increasingly posting reviews about their quality of care, rating both hospitals and physicians on social media platforms (Pai and Alathur, 2018). The increase of digital movements on social media has enhanced patient autonomy and self-determination in medical treatment, underscoring the crucial role of online patient experiences in assessing healthcare quality (Jeyaraman et al., 2023). These ratings are the significant difference among healthcare organizations and providers, and they could significantly impact patients' future behaviour and decisions (Rozenblum & Bates, 2013). Online ratings of patient care are positively correlated with various traditional service and quality metrics in healthcare settings. (Greaves et al., 2013). Several research has utilized online media to understand patient feedback in hospital settings; however, interpreting the large volume of online discussions remains a challenging task (Lagu et al., 2016; Rastegar-Mojarad et al., 2015). Although many papers (Hawkins et al., 2016; Ranard et al., 2016) have explored the use of social media in hospital settings, most focus on using Twitter or Yelp to evaluate hospital service quality, rather than the Facebook platform. For example, since most online posts do not directly address service quality, identifying the most critical posts can be difficult. Hawkins et al. (2016) found that only 10% of hospital-related tweets pertained to service quality. In addition to numerical ratings, the textual content in patient reviews can be a valuable resource for healthcare providers to enhance their services (Rastegar-Mojarad et al., 2015). For example, data on patient experiences is becoming an essential component of the value-based purchasing program proposed by the Centres for Medicare and Medicaid Services (CMS). Finally, Rozenblum and Bates (2013) emphasize the

increasing importance of social media feedback regarding patients' experiences, asserting that this information will complement traditional patient surveys and help identify poor versus outstanding care.

3. Data and methodology

This paper aims to analyse the distribution of customer satisfaction across the main hospitals in the different provinces of the Campania region. We create a database containing 8.024 Google reviews on the main hospitals in Campania by exploiting the Python programming language. The dataset constructed is composed of the following variables: DMO (hospital medical direction), text (review content), date (review date), rate (Google star rating system), and latitude and longitude of the hospital. From the latter, all textless reviews were excluded. Subsequently, sentiment analysis is affected by the reviews. Sentiment analysis is a widely recognized text classification method that measures the direction and/or intensity of the emotional content within a text (Gonçalves et al., 2013). This technique has been utilized across various text analytics fields. For example, in the healthcare sector, sentiment analysis has been used to assess the emotional state of patients in medical or hospital environments (Rodrigues et al., 2016). Specifically, we compare three sentiment analysis models to understand which performed better. These were Feel-it, udpipe, and SentimentR;

- FEEL-IT is an innovative corpus of Italian Twitter posts designed for sentiment analysis. It combines four basic emotions (anger, fear, joy, and sadness) to perform sentiment classification. Created to fill the gap in Italian language resources, FEEL-IT provides broad topical and domain coverage. The corpus has been used to train sentiment recognition models, achieving competitive results on benchmark datasets (Bianchi et al., 2021).
- Udpipe: The udpipe package is a trainable pipeline performing sentence segmentation, tokenization, POS tagging, lemmatization, and dependency parsing. It uses the Universal Dependencies project (UD) as its main information source, whose goal is a multi-language relation in morphology and semantics. Currently, this project has 70 dependency trees on 50 different languages (Wijffels, 2021).
- SentimentR: The sentimentR package is designed to quickly calculate the polarity of sentiment that exists in a text at the sentence level, as well as allowing aggregation by rows or by grouping variables. The goal of sentimentR is to try to strike a balance between efficiency and effectiveness (Rinker, 2021).

For choosing the best model of sentiment analysis, we define the number of false negatives and false positives. The proposed comparison is focused on the combination of sentiment score values and rating values. We define as false negatives all reviews with a positive sentiment score but a negative rating, and conversely, as false positives, those with a negative sentiment score but a positive rating. Table 1 presents the results related to the comparison of three models.

	Udpipe	Feel-it	Sentiment R
False Negatives	26,7%	11%	12,3%
False Positive	4%	2,2%	5,7%

Table 1 – Comparison of three models of sentiment analysis

According to the results in Table 1, we choose Feel-It for performing sentiment analysis. After removing both false negatives and false positives from the dataset, the total number of reviews was reduced to 6,237. Once the sentiment analysis was performed, the sentiment value

was defined through the product between the sentiment score and the rating expressed by the users. To balance the positive and negative ratings, we decided to recode the rating values, as in Table 2, to give equal importance to the extremely positive (5) and extremely negative (1) reviews. This approach stems from the need to define a single value summarizing customer satisfaction for each of the principals considered (Kurniawati, 2023).

Rate	Adjusted rate
1	3
2	2
3	1
4	2
5	3

Table 2 - Adjusted Rating

Finally, to offer a comprehensive overview of the distribution of user sentiment, we created a geolocalized representation. This representation utilizes the aggregate values from each of the five provinces in the Campania region, calculated through an arithmetic mean. We provide a clear visualization of how user sentiment varies across the different areas within the region. This approach allows for an easy comparison and understanding of the sentiment landscape in Campania, highlighting regional differences.

4. Results

To achieve the objectives of measuring customer satisfaction among the primary hospital principals and analyzing its distribution, a geo-referenced map (Figure 1) was utilized. This map represents the aggregate sentiment score, calculated as the average, for each province in the Campania region. The sentiment score for each province is represented in Table 3. It is important to note that the range of variation was adjusted for ease of interpretation. Specifically, it was transformed from the original interval of -3 to +3 to a normalized range of -1 to +1.

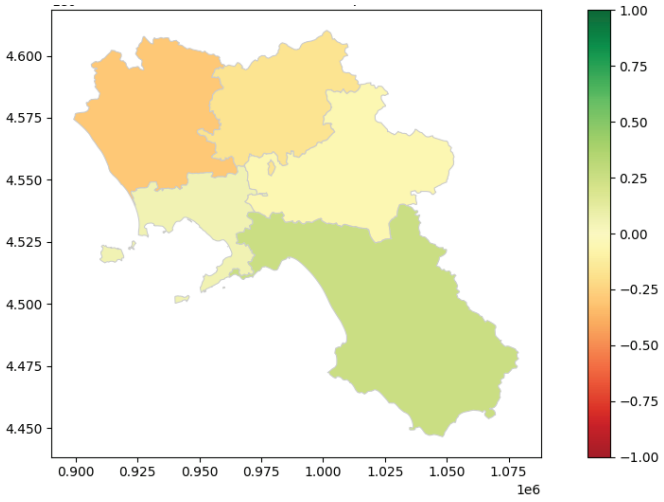


Figure 1 – Sentiment scores by province

Province	Normalized average
Caserta	-0.297024
Benevento	-0.178298
Napoli	0.061799
Avellino	-0.061480
Salerno	0.250844

Table 3 - Sentiment scores normalized and aggregated by province

The map in Figure 1 has two polarities: the most intense green is associated with maximum positivity of reviews, with a maximum value of +1, while red is associated with negative customer satisfaction and has a limit value of -1.

As can be seen from the map and aggregate sentiment score, Naples presents a sentiment score very close to 0, consequently, it is possible to argue that summarily, the positive and negative reviews balance each other out. The same dynamic holds for the Province of Avellino, whose score of -0.061 indicates an internal balance among the various hospital takes. A different discussion is necessary for the remaining provinces. The province of Salerno stands out as the best of the provinces in Campania, with a score of 0.25. It indicates that, as far as Campania is concerned, the most satisfied patients are those who go to the province of Salerno.

Benevento and Caserta, on the other hand, are the two provinces that have the highest negative sentiment score. In these two cases, patient satisfaction does not balance internally. However, Caserta remains the province that has the most dissatisfied patients with their care with a sentiment score of - 0.29. For example, we reported a positive review as well "*Operato presso il reparto di urologia e i medici sono stati bravissimi e non mi hanno mai lasciato solo. Il reparto funziona benissimo con personale professionale valido e con una coordinazione perfetta da parte del coordinatore",* and a negative review as example *"Una cosa vergognosa, totale assenza di organizzazione"*

5. Discussion and conclusion

This study provides a preliminary overview of the sentiment scores derived from Google reviews of major hospitals in the Campania region. By analysing these scores, we have gained valuable insights into the public perception of healthcare services within this area. By utilizing the Feel-It sentiment analysis model (Bianchi et al., 2021), we enhanced the accuracy of sentiment scoring, effectively minimizing false positives and false negatives in the dataset. The geo-referenced sentiment map generated through this analysis provides valuable insights into the distribution of customer satisfaction across the provinces in the Campania region, offering a foundational perspective on patient sentiment toward hospital services. Furthermore, our findings contribute to the broader discourse on healthcare service evaluation by highlighting the utility of sentiment analysis as a robust tool for assessing patient satisfaction (Bovonratwet et al., 2021). Our study addresses critical questions, illustrating how public sentiment can be systematically quantified and analysed to reflect the perceived quality of healthcare services. Additionally, it identifies specific areas for targeted improvement within the hospital system of the Campania region. This approach not only advances methodological frameworks for evaluating healthcare services but also provides actionable insights for policymakers and healthcare administrators seeking to enhance patient experiences. Our analysis has highlighted the necessity of developing a specialized vocabulary to effectively apply textual analysis techniques within the healthcare sector. Establishing such a vocabulary is crucial for ensuring accurate and meaningful analysis of patient feedback and other health-related texts, facilitating better understanding and improving healthcare services. This aligns with the findings of

Eichstaedt et al. (2021), who emphasized the importance of domain-specific lexicons in sentiment analysis for achieving more precise results. For future developments, we aim to extend our analysis to include the entire Italian territory, providing a comprehensive evaluation of all healthcare facilities across the country. This expanded scope will enable us to move beyond a regional perspective, such as our current focus on Campania, and allow for a nationwide assessment of hospital services. By analysing sentiment scores and customer satisfaction at a national level, we can identify overarching trends and disparities between regions, gaining a more holistic understanding of public perceptions regarding healthcare services. Such a broader approach will also facilitate comparisons between regional and national dynamics, helping to contextualize our findings and offering valuable insights into the overall quality of healthcare in Italy. Ultimately, the nationwide analysis will contribute to a more informed discussion on how to address systemic issues, improve patient experiences, and enhance the effectiveness of hospital services throughout the country.

References

Bardach, N. S., Asteria-Peñaloza, R., Boscardin, W. J., Dudley, R. A. (2013). The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Quality & Safety*, **22**(3), pp. 194-202.

Bianchi, F., Nozza, D., Hovy, D. (2021). 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*. *Association for Computational Linguistics*.

Bovonratwet, P., Shen, T. S., Islam, W., Ast, M. P., Haas, S. B., Su, E. P. (2021). Natural language processing of patient-experience comments after primary total knee arthroplasty. *The Journal of Arthroplasty*, *36*(3), pp. 927-934.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., ...Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, **26**(4), 398.

Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the First ACM Conference on Online Social Networks*, pp. 27-38.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L. (2013). Harnessing the cloud of patient experience: Using social media to detect poor quality healthcare. *BMJ Quality & Safety*, **22**(3), pp. 251-255.

Hariri, R. H., Fredericks, E. M., Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, **6**(1), pp. 1-16.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., Greaves, F. (2016). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety*, **25**(6), pp. 404-413.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... Greaves, F. (2016). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety*, *25*(6), pp. 404-413.

Hopper, A. M., Uriyo, M. (2015). Using sentiment analysis to review patient satisfaction data located on the internet. *Journal of Health Organization and Management*, **29**(2), pp. 221-233.

Jeyaraman, M., Ramasubramanian, S., Kumar, S., Jeyaraman, N., Selvaraj, P., Nallakumarasamy, A., Yadav, S. (2023). Multifaceted role of social media in healthcare: opportunities, challenges, and the need for quality control. *Cureus*, **15**(5).

Kurniawati, P. (2023). Sentiment analysis of maxim online transportation app reviews using support vector machine (svm) algorithm. *Building of Informatics Technology and Science*, **5**(2).

Lagu, T., Goff, S. L., Craft, B., Calcasola, S., Benjamin, E. M., Priya, A., Lindenauer, P. K. (2016). Can social media be used as a hospital quality improvement tool? *Journal of Hospital Medicine*, **11**(1), pp. 52–55.

Liao, H., Qi, J., Zhang, J., Zhang, C., Liu, F., Ding, W. (2024). Mining and fusing unstructured online reviews and structured public index data for hospital selection. *Information Fusion*, **103**, 102142.

Pai, R. R.,Alathur, S. (2018). Assessing mobile health applications with twitter analytics. *International Journal of Medical Informatics*, **113**, pp.72-84.

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., ... Merchant, R. M. (2016). Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs*, **35**(4), pp. 697-705.

Rastegar-Mojarad, M., Ye, Z., Wall, D., Murali, N., Lin, S. (2015). Collecting and analyzing patient experiences of health care from social media. *JMIR Research Protocols*, **4**(3).

Rinker, T. W. (2021). sentimentr: Calculate text polarity sentiment (Version 2.9. 1)[Computer software].

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., Rosa, T. C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, **85**(1), pp. 80–95.

Rozenblum, R., & Bates, D. W. (2013). Patient-centred healthcare, social media and the internet: The perfect storm?. *BMJ Quality & Safety*, **22**(3), pp. 183-186.

Wijffels, J. (2021). udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the'UDPipe''NLP' toolkit. 2020. *R Package Version* 0.8, 8.

Zaman, N., Goldberg, D. M., Abrahams, A. S., Essig, R. A. (2021). Facebook hospital reviews: Automated service quality detection and relationships with patient satisfaction. *Decision Sciences*, **52**(6), pp. 1403-1431.

doi.org/10.26398/asaproc.00127

Markov switching autoregressive models and Bayesian hydrology

Roberta Paroli[a], Luigi Spezia[b]
[a]Department of Statistical Science, Università Cattolica SC, Milano, Italy.
[b]Biomathematics & Statistics Scotland, Aberdeen, UK

1. Introduction

Hydrological time series are realizations of complex stochastic systems. A few issues need to be taken into account by the modellers: non-Normality, non-linearity, non-stationarity, and long memory. Non-normality is observed when the data density is multimodal or asymmetric or kurtic and the data cannot be considered as realisations from a Gaussian process. Non-linearity is assumed when the whole series does not show the same statistical peculiarities over all the observations, but they can be classified into a few homogeneous groups, each one with specific characteristics (e.g., different means and/or different variances). Non-linearity can also be assumed when the series exhibits asymmetries, e.g., when peaks are sharper (or more rounded) than the troughs, and/or when the cycles increase at a different rate from which they decrease. Weak non-stationarity is caused by generating processes having time-varying means and autocovariances (possibly due to periodic components and/or covariates). Finally, when the series shows high autocorrelations at the higher lags, with slow decay, the observations are realizations from a long memory process.

Because of these issues, hydrological time series can be effectively analysed by Markov switching autoregressive models (MSARMs). MSARMs are pairs of discrete-time stochastic processes, one observed and one latent, or hidden. The hidden process is a finite-state Markov chain, whereas the observed process, given the Markov chain, is conditionally autoregressive. The dynamics of the observed process is driven by the dynamics of the latent one, so that each observation depends on the contemporary state of the Markov chain.

We propose MSARMs within the Bayesian framework: Bayesian inference, model choice, and variable selection are performed numerically by Markov chain Monte Carlo algorithms.

Our applications to isotope signatures, turbidity measurements, and river temperatures provide clear examples of the suitability of the MSARMs in hydrological time series analysis in particular and environmental sciences in general.

2. Bayesian Markov switching autoregressive models

Markov Switching Autoregressive Models (MSARMs) are discrete-time stochastic processes $(\{Y_t\},\{X_t\})$, such that $\{Y_t\}$ is a sequence of observed events on the real line, whereas $\{X_t\}$ is an unobserved finite-state Markov chain, hidden behind the observed process. The sequence $\{Y_t\}$ given $\{X_t\}$, must satisfy both the order-p dependence and the contemporary dependence condition: each random variable Y_t ($t=p+1,\dots,T$) depends on the p previous observations $Y_{t-1},\dots,Y_{t-p}$, and its conditional distribution, given the whole sequence of hidden states, depends only on the contemporary hidden state, or regime, X_t.

Let $S_X=\{1,\dots,m\}$ be the state space of the hidden Markov chain, where $\Gamma=[\gamma_{j,i}]$ is the $m\times m$ transition matrix, with $\gamma_{j,i}=P(X_t=i|\, X_{t-1}=j)$, $0<\gamma_{j,i}<1$, for any $i,j\in S_X$ and any $t=p+1,\dots,T$. When $x_t=i$ ($t=p+1,\dots,T$; $i=1,\dots,m$), the equation describing the MSARM of order $(m;p)$, henceforth MSAR$(m;p)$, is

$$y_t = \mu_i + \sum_{\tau=1}^{p} \varphi_{i,\tau} y_{t-\tau} + \varepsilon_t,$$

where $\{\varepsilon_t\}$, given $\{x_t\}$, is a Normal white noise process of precision (the reciprocal of variance) λ_i,

with μ the vector of the m intercepts μ_i; $\varphi=[\varphi_{i,\tau}]$ the matrix of the mxp autoregressive coefficients ($\tau=1,\ldots,p$); and λ the vector of the m precisions λ_i.

If some dynamic covariates or periodic components are also available, they can be included in the model, both in the observed and the hidden process. Let $z_{t,k}$ be q state-dependent exogenous variables; η_t a harmonic component and $\beta_{t,i}$ another harmonic component depending on the current state i of the Markov chain which are assumed to reflect any seasonality in the time series. The general formulation of the model is:

$$y_t = \mu_i + \sum_{\tau=1}^{p} \varphi_{i,\tau} y_{t-\tau} + \sum_{k=1}^{q} \theta_{i,k} z_{t-1,k} + \eta_t + \beta_{i,t} + \varepsilon_t,$$

with $\theta=[\theta_{i,k}]$ the matrix of the mxq covariate coefficients ($i=1,\ldots,m$; $k=1,\ldots,q$).

The first harmonic component η_t has periodicity $2s$:

$$\eta_t = \sum_{j=1}^{s^*} \eta_{1,j} \cos\left(\frac{\pi j t}{s}\right) + \eta_{2j} \sin\left(\frac{\pi j t}{s}\right),$$

where s^* is the number of relevant harmonics ($s^* \leq s$). The second harmonic component $\beta_{t,i}$ has periodicity $2r$ and depends on the current state i of the Markov chain:

$$\beta_t = \sum_{j=1}^{r^*} \beta_{i,1,j} \cos\left(\frac{\pi j t}{r}\right) + \beta_{i,2,j} \sin\left(\frac{\pi j t}{r}\right),$$

where r^* is the number of relevant harmonics ($r^* \leq r$). In the case of an annual periodicity $2s=365$, whereas for a daily cycle $2r=24$.

On the other hand, the hidden Markov chain can be generalized by assuming the transition probabilities are time-varying and modelled as logistic functions of the covariates:

$$\Gamma^t=[\gamma^t_{j,i}] \qquad \gamma^t_{j,i} = P(X_t=i \mid X_{t-1}=j),$$

with

$$logit\left(\gamma^t_{j,i}\right) = \ln\left(\frac{\gamma^t_{j,i}}{\gamma^t_{j,j}}\right) = \sum_{k=0}^{q} \alpha_{j,i,k} z_{t-1,k}$$

where $z_t=(1,z_{t,1},\ldots,z_{t,q})'$, for any $t=p,\ldots,T-1$, and $\alpha=[\alpha_{i,j}]$ the matrix of the coefficient vectors, i.e. $\alpha_{j,i} = \left(\alpha_{j,i,0}, \alpha_{j,i,1}, \ldots, \alpha_{j,i,q}\right)$ if $i \neq j$ or equal to the null vector if $i = j$.

Some of the q state-dependent exogenous variables affecting the observed process might be redundant according to the state visited by the hidden Markov chain. Redundancies in the covariates might also be present in those appearing in the time-varying transition probabilities. A selection of the most relevant explanatory variables is recommended for a better interpretation of both the observed and the latent phenomenon. In order to perform a stochastic selection within the different sets of covariates, each exogenous variable can be associated with a binary indicator: when the indicator equals one the corresponding variable is included in the model, otherwise the covariate is excluded.

The vector of the unknown parameters of a MSAR($m;p$) is ($\mu, \lambda, \varphi, \theta, \eta, \beta, \alpha, x^T, y_{miss}$), where $x^T=(x_{p+1},\ldots,x_T)$ is the sequence of hidden states; and y_{miss} the set of all missing values occurring within the observed data.

To complete the model, we set independent fairly vague priors, invariant to the relabelling of the hidden states. When we have m states, we have $m!$ ways to label them; when the priors are invariant to the different state permutation, $m!$ different models are interchangeable by permuting their labelling. This is the so-called label switching problem (see, among others, Frühwirth-Schnatter, 2001; Spezia, 2009) and it can be overcome by placing some parameters in increasing or decreasing order. Experience (Spezia et al., 2004; Paroli and Spezia, 2008a) suggests that in environmental time series hidden states can be ordered by increasing levels of variability.

Bayesian inference is then performed using appropriate MCMC algorithms along three consecutive steps: *i*) variable selection (by means of the MKMK algorithm, see Paroli and Spezia 2008b); *ii*) model choice (by means of Bayes factor, where the natural logarithms of the marginal likelihoods are estimated numerically), ; *iii*) parameter estimation, model fitting analysis, missing values restoration, hidden states reconstruction and observations classification into regimes.

In the following sections we will briefly describe three applications of MSARMs to

hydrological time series analysed in joint work by the Authors in previous papers.

3. Application 1: Isotope signatures

The first application (Birkel et al., 2012) is related to the analysis of the temporal dynamics of the time series of two conservative water isotopes, deuterium (δ2H) and oxygen-18 (δ18O), in daily stream water samples, over two years (from May 1st 2008 to April 30th 2010), in a small catchment in eastern Scotland. The hidden states of the Markov chain could be associated with meteorological and hydrological drivers identifying the short (event) and longer-term (inter-event) transport mechanisms for both isotopes. They represent the unobserved levels of variability in the observed process that result from complex interactions of meteorological and hydrological conditions. Meteorological and hydrological time series are also used as covariates in the autoregressive structure of the model along with a yearly periodic component to identify seasonal components. Even though the memory retained by the chemical composition of streamflows from past chemical disturbances is long-lasting (e.g., expressed by the long tail of travel time distribution models), we considered contemporary covariates for model selection only. Eight available covariates are considered as potentially explanatory of the dynamics of both isotopes: air temperature, turbidity, relative humidity, discharge, net radiation, evapotranspiration rate, precipitation volume and both isotopes in precipitation.

The result of model choice is a MSARM with 2 hidden states and autoregression of order 1, i.e. MSAR(*2;1*) as the best model for oxygen-18, whereas the best model for deuterium is a MSAR(*2;2*). The covariates with the strongest influence on the stream isotope time series are selected by the MKMK variable selection algorithm. Both isotopes seem to require air temperature as an important explanatory variable, presumably reflecting the annual periodicity (seasonality). Rainfall input signatures of both isotopes were selected for the MSARMs driving especially the short-term response in the stream. Precipitation volume, however, has only been selected for the deuterium model and the turbidity time series showed much more explanatory power than, for example, discharge or other climatic variables for both isotopes. This is also assumed to be related to the short-term dynamics of turbidity, which may have a similar underlying driver to the short-term isotope response even for relatively small runoff events reflecting quick near-surface flow pathways. The model fit is satisfactory, as shown by comparing observed and fitted values in Figure 1. The trends of the series are well described. Figure 1 also includes 95% credible intervals: four oxygen-18 observations lie outwith the interval, whereas eight deuterium observations are not bracketed by the intervals (1 above and 7 below). Extreme events cannot be captured by the model, as expected, but the fitting of the series is also made more difficult by the occurrence of the 17% missing values.

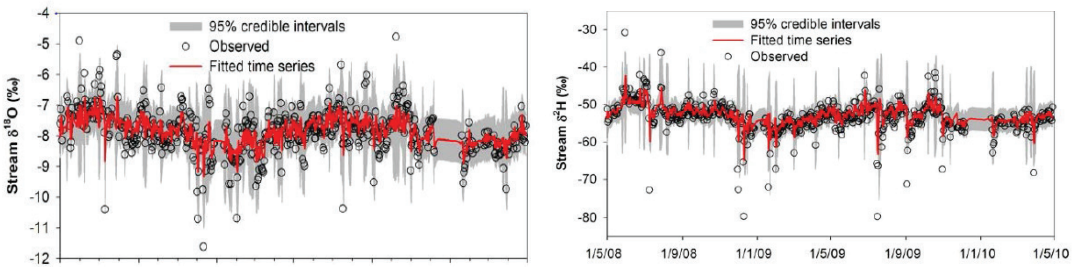


Figure 1: observed (black) and fitted (red) values with 95% credible intervals

Finally, the observations are split into two groups, according to the corresponding hidden state: state 1 underlies the group of observations with the highest variability, whereas state 2 underlies that with the lowest variability. The states effectively separate short (state 1) and longer-term (state 2) controls on the isotope dynamics. Results show that the oxygen-18 series was influenced by short-term (event) dynamics in form of isotopic precipitation input for 17% of the time but only 9% of the time for deuterium. State 1 effectively expresses the greatest explanatory power of rainfall

isotope signatures on the stream response corresponding to storm event conditions and the non-linear activation of quick near-surface flow paths.

4. Application 2: Turbidity measurements

A turbidity time series, recorded every 15 minutes in a first order Scottish stream (the Wemyss catchment) for more than a year (from 1st January 2011 to 5th January 2012), along with two covariates (stage height and rainfall), is considered in Spezia et al. (2021).

Environmental works investigate the use of continuous automated turbidity as a proxy to estimate particulate phosphorus, fine sediment or hydrophobic pollutant, loads using site specific calibrations of turbidity versus the pollutant of interest, with some success. However, the approach is limited by only moderately good correlations, and therefore methods that enhance the fit, for example between turbidity and phosphorus loads, are needed. The turbidity trace along with discharge and other data may contain hidden temporal patterns (Birkel et al., 2012), that can be used to improve these relationships and so improve estimates of pollutant and fine sediment loads and understanding of sources and processes delivering them to surface waters. Turbidity time series have complex dynamics because they are non-linear, non-Normal, non-stationary, with a long memory, and present missing values. The best model chosen by our algorithm (computing the logarithms of the marginal likelihood by Evolutionary Monte Carlo via the power posteriors) is a MSAR(*3,4*), that is characterized by three hidden states (*m=3*) and autoregressions of the fourth order (*p=4*). From the results of the parameter estimation, it can be seen that the covariates in the observed process have a positive coefficient, that is the level of turbidity increases when stage height and/or rainfall increase. On the other hand, the covariates in the hidden process have a negative coefficient, that is the probabilities of state transitions decrease when stage height and/or rainfall increase, while the diagonal probabilities of the transition matrices increase. These permanence probabilities decrease as the labels of the state increase. As a consequence, the corresponding hypothetical mean time spent in a state decreases as well: 1.95, 0.65, and 0.59 hours.

We could also reconstruct the sequence of the hidden states by computing the posterior mode over the EMC sample. The hidden Markov chain visited states 1, 2, and 3 a total of 23,103 times, 8,715 times, and 3,664 times, respectively. Given the latent regimes, the observations were classified into the three hidden states with increasing levels of variability (see Figure 2).

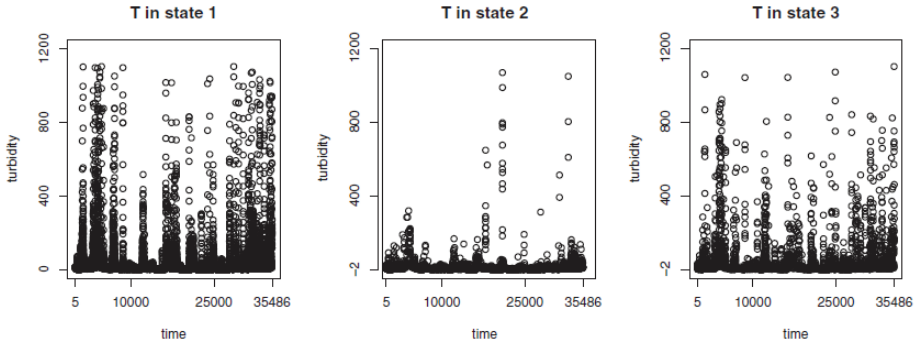


Figure 2: observations classified into three hidden states with increasing variability

Both the observed and the unobserved process at the current time were depending on covariates at the previous time. Thus, given the current latent regimes, the previous stage height and rainfall were classified into the three hidden states as well. The model fit was very satisfactory, as shown by comparing actual and fitted values (Figure 3). The model performance was assessed by the RMSE and the MAE, which are very low. All observations were within the 99% credibility interval.

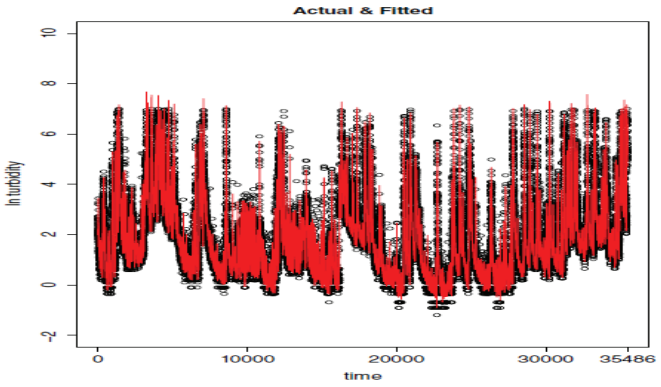


Figure 3: comparison between actual (black) and fitted (red) values

5. Application 3: River temperatures

River temperature and its temporal variability at sub-daily to seasonal scales influence in-stream biochemical processes and aquatic ecology (e.g., the life cycle and habitat of salmonid species). The dominant controls on river temperature and its variations are thermal inputs into the river system and the entire catchment. Further influences include hydrology (e.g., stream-flow volume, snowmelt, base-flow contribution) and channel and landscape characteristics such as riparian vegetation. Information on the temporal variability of river temperature and its drivers is important for understanding temperature-dependent environmental processes but also for environmental management and planning. Therefore, modelling approaches are applied to estimate river temperatures under potential future climate and land cover changes, to reconstruct past river temperature and to extend the spatial coverage of river temperature observations by physical or statistical models.

In Spezia et al. (2023) the dynamic variability of water temperature is investigated by analysing an hourly water temperature time series automatically recorded in the Gairn catchment (Scotland) for more than five years (from 16th August 2012 to 23rd November 2017). Seven covariates are also incorporated into the model through both the hidden Markov chain and the observed process. The covariates are river flow, air temperature, rainfall, wind speed, wind direction, radiation, and soil temperature. The data set is also characterised by periodicities: the hourly temperatures vary according to the dynamics of the year and of the 24 hours of the day. Hence, both an annual and a state-dependent daily harmonic component are added to the observed process.

In this work, we consider only series of covariates with no missing values. Thus, we handle 50-day long short series when working with seven covariates; 810-day long intermediate series with three covariates; and a 1926-day long whole series with a single covariate. This because we do not want that our analyses are biased by the missing values within the exogenous variables, so three applications are considered.

The results show that for the longest series with a single covariate, the best model has three hidden states and autoregressions of the fifth order (MSAR(*3,5*)). Flow is relevant in the observed process for two states only, while it is not selected in the hidden process and the Markov chain is homogeneous. For the intermediate series with three covariates, the best model has three hidden states and autoregressions of the sixth order (MSAR(*3,6*)). Flow is relevant in the observed process for one state only, while air temperature was always selected both in the observed and the hidden process (the Markov chain is non-homogeneous and the transition probabilities are time-varying). For the short series with seven covariates, we obtained that the best model is the linear autoregression of the sixth order, without a hidden Markov chain (MSAR(*1,6*)). Air temperature, solar radiation, and soil temperature are the relevant variables to explain the water temperature dynamics. Thus, discharge is a proxy for water temperature modelling, when no other more directly related variables are available. In those situations, the latent states will help to model the long-term dynamics, in the absence of true predictors with a physical meaning.

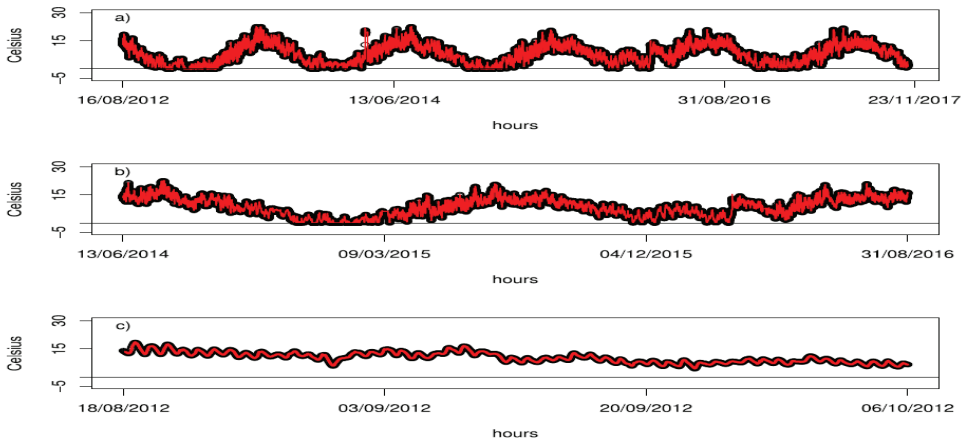


Figure 4: comparison between actual (black) and fitted (red) values

The model fit is very satisfactory, as shown by the comparison of actual and fitted values (Figure 4). The RMSE and MAE are very low and all observations were within the 99% credibility interval. As we saw in the first two applications, the hidden regimes can have an interpretation related to the seasonality. In fact, the Markov chain shows an annual dynamic which anticipates the annual dynamics of the water temperatures. The observations in the decreasing part and the bottom of the cycle are those with the lowest variability and associated with state 1; those in the increasing part, the top, and the beginning of the decreasing part of the cycle have a higher level of variability and are associated with state 2; finally, in state 3 we have the observations with the highest level of variability and they are mostly in the increasing part of the cycle. It is not surprising that for the short series (50 days, i.e. no annual periodicity) the model is not multi-state. It would be interesting to see what happens when considering the seven covariates on a longer series, that is if the same covariates are selected in a non-linear model ($m>1$).

References

Birkel, C., Paroli, R., Spezia, L., Dunn, S.M., Tetzlaff, D., Soulsby, C. (2012). A new approach to simulating stream isotope dynamics using Markov switching autoregressive models. *Advances in Water Resources*, **26**, pp. 308-316.

Birkel, C., Soulsby, C., Tetzlaff, D., Dunn, S.M., Spezia, L. (2012). High-frequency storm event isotope sampling reveals time-variant transit time distributions and influence of diurnal cycles. *Hydrolological Processes,* **26**, pp. 308-316.

Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of American Statistical Association,* **96,** pp. 194-209.

Paroli, R., Spezia, L. (2008a). Bayesian inference in non-homogeneous Markov mixture of periodic autoregressions with state-dependent exogenous variables. *Computational Statistics and Data Analysis,* **52,** pp. 2311-2330.

Paroli, R., Spezia, L. (2008b). Bayesian variable selection in Markov mixture models. *Communication in Statistics – Simulation and Computation,* **37,** pp. 25-47.

Spezia, L., Futter, M.N., Brewer, M.J. (2011). Periodic multivariate normal hidden Markov models for the analysis of water quality time series. *Environmetrics,* **22,** pp. 304-317.

Spezia, L., Gibbs, S., Glendell, M., Helliwell, R., Paroli, R., Pohle, I. (2023). Bayesian analysis of high-frequency water temperature time series through Markov switching autoregressive models. *Environmental Modelling & Software,* **167**, 105751.

Spezia, L., Vinten, A., Paroli, P., Stutter, M. (2021). An evolutionary Monte Carlo method for the analysis of turbidity high-frequency time series through Markov switching autoregressive models. *Environmetrics*, **32**, e2695.

doi.org/10.26398/asaproc.00128

Collecting ADS-B data for timely production of air traffic statistics

Samanta Pietropaoli, Massimiliano Amarone
Italian National Institute of Statistics, Rome, Italy.

1. Introduction

The aviation industry plays a pivotal role in global transportation and commerce, facilitating the movement of millions of passengers and vast quantities of goods across continents daily. Accurate and timely analysis of flight transport data is crucial for optimising operational efficiency, enhancing safety measures, managing air traffic, and supporting strategic decision-making processes within the industry. Traditionally, data collection in this domain has relied on official sources, such as aviation authorities and airlines, which report data to national statistics agencies. For instance, the National Institute of Statistics conducts the Air Transport Survey under Regulation (EC) No. 437/2003. This survey aims to gather detailed statistical information on aircraft movements at Italian airports, air routes, and the quantities of passengers, goods, and mail transported. While these official sources provide essential data, they often face limitations related to data availability, timeliness, and granularity. Specifically, the reliance on manual data submission processes can result in delays, typically with a lag of at least one month, and may lack the real-time granularity required for dynamic operational decision-making.

This study explores the efficacy of automatic dependent surveillance-broadcast (ADS-B) data as a complementary approach to conventional data collection methods for obtaining and analysing flight transport data. By integrating ADS-B data with official statistics, this research aims to enhance the comprehensiveness, accuracy, and timeliness of air traffic data analysis. The potential benefits of utilising ADS-B data include real-time monitoring capabilities, improved data granularity, and reduced reliance on manual data reporting processes. Additionally, the study investigates the application of machine learning algorithms to estimate passenger numbers and cargo volumes from ADS-B data, addressing the inherent limitations of ADS-B in directly capturing these metrics.

The objectives of this study are to compare the effectiveness of ADS-B data against traditional data sources in capturing flight movements and associated cargo information and to develop and evaluate machine-learning models for estimating passenger and cargo volumes. By addressing these objectives, the study aims to demonstrate the value of integrating ADS-B data into existing data collection frameworks, thereby contributing to more robust and real-time air traffic statistics.

2. The ADS-B data and the survey data: Characteristics and differences

Automatic dependent surveillance-broadcast (ADS-B) is a communication protocol that facilitates the exchange of flight information, including speed and location, between airplanes and air traffic control centers. An aircraft determines its position and speed using global navigation satellite systems and then broadcasts this information, along with other details like emergency status and flight number, as ADS-B out signals (Karam, 2022).

ADS-B enables real-time tracking by periodically broadcasting an aircraft's position. This data is publicly accessible and can be sourced from platforms like FlightAware, which aggregates and provides crowdsourced ADS-B data. The availability of this data is real-time, and downloads from the FlightAware APIs are automatically managed through informatics procedures. Conversely, the Istat survey is comprehensive, requiring all Italian airport management companies to provide data. Participation in the survey is mandatory, with all requested information needing to be submitted to Istat by the 30th day of the month following the

reference month for the data. This results in a data availability delay of at least one month. Therefore, the use of ADS-B technology presents a significant opportunity to obtain exhaustive real-time data, complementing the delayed data from the Istat survey.

The survey focuses on the monthly movements at national airports by both Italian and foreign aircraft, along with the associated cargo transported, including passengers, goods, and mail. Therefore, ADS-B data does not directly provide statistics on loaded/unloaded goods and mail, boarded/disembarked passengers, direct transit passengers, and available seats for passenger transport, but it is possible to estimate these metrics based on the type of aircraft and the last time series data[1]. This estimation process, combined with the real-time nature of ADS-B data, allows for the production of flash indicators on tourism and trade.

3. Collecting ADS-B data with the APIs

To collect ADSB data, we utilised the highly customizable FlightAware APIs. These APIs allow us to integrate accurate real-time, historical, and predictive flight insights across all segments of the aviation industry into our products and applications (Martín Coll, 2018). These APIs provide live positional data, flight details (including schedules and gate information), block times, autopilot modes, and weather reports (e.g., wind speed/direction, air pressure). The FlightAware data acquisition workflow is designed to be efficient and automated.

Data is acquired through URL requests to an API, with a Java application consisting of three programs: *DownloadJsonData* (downloads JSON data), *UploadJsonData* (uploads data to the database), and *SendMailifErrorFound* (sends error reports). The application runs daily on a Linux server via crontab, ensuring automatic updates. The *DownloadJsonData* program composes an HTTP client request, setting the API key for service authorisation in the request header and including the date and airport of departure or arrival in the URL. Upon executing the request, the code processes the JSON response by mapping it to Java classes and saving the file on the server. In case of service unavailability[2], the program manages waiting periods of 5-10 minutes before making further attempts. The *UploadJsonData* program reads all JSON files related to both departures and arrivals that were previously downloaded, maps them to specific Java classes, and inserts the contained data into the database in bulk (batch mode). The *SendMailIfErrorFound* program reads the log files of the previous programs to identify any warnings or errors and sends the relevant notifications to a list of recipients. These recipients will then take action to identify the issue and execute emergency programs for data recovery. The FlightAware service for arrivals and departures requests offers a 10-day history from the current date of the request, which is useful for data recovery in case of service unavailability. This comprehensive approach ensures reliable and efficient air traffic data acquisition and management, leveraging robust Java applications and the extensive capabilities of the FlightAware APIs.

4. Estimation of passengers and goods

As previously mentioned, the data downloaded from FlightAware, compared to the data obtained from Official Statistics, do not provide information regarding passengers and cargo transported. Therefore, it is necessary to train a machine-learning model that predicts these values based on certain features. In our case study, we focused our analysis on flights arriving at and departing from eight major Italian airports, distributed across the territory as follows: one in the north-west (Milan Malpensa), two in the north-east (Venice and Bologna), two in

[1] For example, the capacity and typical load factors of specific aircraft models can be used to approximate the number of passengers and the volume of goods transported.

[2] e.g., 500 Internal Server Error, 502 Bad Gateway, 503 Service Unavailable

the center (Rome Fiumicino and Pisa), two in the south (Naples and Bari), and one on the islands (Catania). The analysis specifically considered data from June 2023[3]. Table 1 presents the number of arriving and departing flights as reported by the official statistics and recorded by the FlightAware APIs, along with the percentage of flights successfully merged between the two sources relative to the official statistics data. The merging process was based on flight details such as the scheduled date, flight code, airline operator, and airport code. The match rates between the two data sources varied by airport and direction (arrivals and departures). Across all airports, match rates ranged from 80% to 91%, reflecting variability in data consistency and coverage between the official statistics and FlightAware APIs.

In our study, we applied two machine-learning algorithms, random forest (RF) and decision tree (DT) regressors, to estimate passenger numbers and cargo volumes using available flight data. We evaluated their performance using cross-validation and the R-squared (R^2) metric. This metric evaluates how well a model fits the data and predicts future outcomes. Specifically, R^2 quantifies the proportion of variance in the actual values that is explained by the predictions[4]. We also calculated 95% confidence intervals (CI) using bootstrap resampling with 10,000 iterations[5].

The choice of models for this analysis was driven by the unique characteristics of air traffic data, which often exhibit complex patterns influenced by seasonality, demand fluctuations, and varying operational conditions. These challenges require models capable of capturing non-linear relationships and interactions between predictors while maintaining robustness against overfitting. Decision Trees were chosen for their interpretability, but they are prone to overfitting, limited out-of-sample performance, and high variability when used independently. These limitations are well-documented in the literature (Breiman, 2017). To address these issues, we incorporated RF, a tree ensemble model, in our analysis. RF aggregate predictions from multiple DT, each trained on a random subset of the data, to improve predictive performance and reduce model variance (Géron, 2019). By introducing randomness during training, RF mitigates overfitting and improves generalization. Additionally, RF are well-suited for datasets with complex, non-linear interactions and are widely recognized as a robust choice for regression and classification tasks (Breiman, 2001). The selection of these models was further motivated by the structure of the flight data, which includes both numerical features (e.g., available seats) and categorical variables (e.g., aircraft type, airport code). DT and RF inherently handle categorical variables without extensive preprocessing, making them practical for this type of analysis. Furthermore, the interpretability of DT and the predictive power of RF align with our dual: understanding the factors influencing air traffic metrics and achieving accurate predictions. By combining these models, we leverage their complementary strengths to comprehensively analysis of passenger numbers across all the eight airports while estimating cargo volumes for a subset of them.

Several features were selected for their relevance to both passenger and cargo estimations. All features were derived from FlightAware data, except for the number of available seats, which was sourced from Official Statistics. The data preparation process also included transforming categorical features into dummy variables before applying the models.

Table 1. Number of flights collected by official statistics in June 2023, recorded by the FlightAware APIs and percentage of flights successfully merged.

Airports	Official	Flight Aware APIs	Merged data

[3] It is worth noting that the data refer to June 2023, a month when the data collection system via FlightAware APIs reached a stable operational phase, resulting in improved data quality and consistency.

[4] The R^2 ranges from 0 to 1, with values closer to 1 indicating a better fit and stronger predictive performance. For example, if the R^2 value is 0.80, it means that 80% of the variance in the data points is explained by the model.

[5] Confidence intervals are a flexible, non-parametric method that avoids distributional assumptions and quantifies uncertainty by providing a range for the true parameter.

		Statistics		n.	%
Bari	Arrivals	2,094	2,166	1,782	85
	Departures	2,093	2,155	1,793	86
Catania	Arrivals	3,397	3,540	2,928	86
	Departures	3,400	3,485	2,969	87
Milan Malpensa	Arrivals	7,228	8,958	6,443	89
	Departures	7,230	8,671	6,443	89
Bologna	Arrivals	3,266	3,537	2,834	87
	Departures	3,258	3,596	2,860	88
Venice	Arrivals	3,706	3,871	3,166	85
	Departures	3,713	3,853	3,189	86
Rome Fiumicino	Arrivals	11,710	12,058	10,622	91
	Departures	11,703	12,007	10,627	91
Naples	Arrivals	4,063	4,240	3,272	80
	Departures	4,073	4,204	3,343	82
Pisa	Arrivals	1,576	1,650	1,307	83
	Departures	1,578	1,640	1,309	83

Source: Authors' elaboration on Istat data and FlightAware data

The selected features, gathered from merged flight datasets for arrivals and departures, included:

- Flight arrival and departure times, recorded by the scheduled hour to distinguish between morning, afternoon, evening, and night flights;
- Day of the week, derived from the scheduled date and expressed as a number (0 for Sunday through 6 for Saturday);
- Aircraft type, which incorporates information about the flight code and airline;
- Total number of available seats, indicating the passenger capacity of the flight;
- Destination (for departures) or origin (for arrivals) airport codes, used to identify the specific flight route.

Table 2 presents the R^2 scores and 95% CI for passenger predictions (arrivals and departures) using RF and DT regressors across eight airports in June 2023[6]. The results showed that, for most airports, the RF model outperforms the DT model, demonstrating higher R^2 scores and narrower confidence intervals. For instance, Milan Malpensa shows a high R^2 of 0.8217 for both arrivals and departures using RF, with a confidence interval of [0.7862, 0.8491], indicating strong predictive accuracy. In contrast, DT regressor gives lower R^2 scores (0.7816 for both arrivals and departures). Similarly, Rome Fiumicino also favors the RF model with an R^2 of 0.8610 for arrivals, significantly higher than the DT's 0.8235. In some airports, like Naples and Bologna, both models perform less effectively. For example, Naples exhibits the lowest R^2 scores, with RF achieving only 0.6875 for arrivals and 0.5662 for departures, while the DT performs even worse. In Bologna, both models yield similar performance, with R^2 scores just over 0.5 for both arrivals and departures, reflecting challenges in predicting passenger numbers at these airports.

For cargo predictions, reliable results for both arrivals and departures are achieved only for Rome Fiumicino and Milan Malpensa. For example, in the case of arrivals, Rome Fiumicino shows R^2 scores of 0.6539 (CI: [0.5458, 0.7130]) for RF and 0.5337 (CI: [0.3157, 0.6348]) for DT, while Milan Malpensa achieves R^2 scores of 0.7364 (CI: [0.6442, 0.8096]) for RF and 0.6896 (CI: [0.6007, 0.7653]) for DT. For other airports, characterized by smaller data volumes as noted in Table 1, the models fail to adapt effectively, suggesting the need to

[6] The datasets were split so that 80% of the data were used as a training-validation set while the remaining 20% was put aside and used as a test set to check the quality of the model obtained, that is its capability to correctly predict the outcome of uncollected passengers and cargo volumes flights.

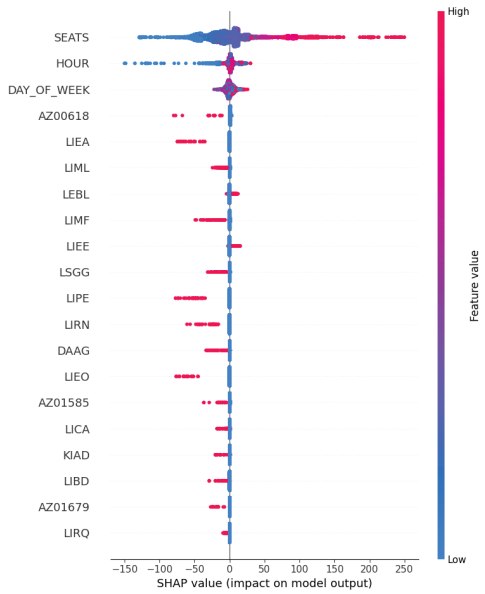
revise the training algorithm using data from multiple months to improve predictive accuracy. In particular, the airports of Pisa and Naples yield the poorest estimates, with R^2 values even falling into negative value. This is likely not coincidental, as both cities are major port hubs where cargo is predominantly transported by sea. These findings prompt a critical reconsideration of the feasibility of using the RF model to estimate cargo volumes across all Italian airports and underscore the need to extend the data period used for training the model.

Table 2. Model accuracy: R^2 score and 95% confidence interval for passengers predictions, for random forest and decision tree regressors (June 2023)

Airports	Regressors	Passengers			
		Arrivals		Departures	
		R^2	CI	R^2	CI
Bari	RF	0.7346	[0.6876, 0.7980]	0.7019	[0.6571, 0.7517]
	DT	0.7224	[0.6831, 0.7958]	0.6784	[0.6271, 0.7328]
Catania	RF	0.7727	[0.7354, 0.8099]	0.7576	[0.7167, 0.7746]
	DT	0.7077	[0.6745, 0.7483]	0.6914	[0.6431, 0.7143]
Milan Malpensa	RF	0.8217	[0.7862, 0.8491]	0.8217	[0.7862, 0.8491]
	DT	0.7816	[0.7431, 0.8083]	0.7816	[0.7431, 0.8083]
Bologna	RF	0.8099	[0.7945, 0.8277]	0.6813	[0.6399, 0.7324]
	DT	0.7554	[0.7246, 0.7836]	0.6061	[0.5737, 0.6377]
Venice	RF	0.7506	[0.7165, 0.7847]	0.7617	[0.7283, 0.8046]
	DT	0.6841	[0.6572, 0.7301]	0.7078	[0.6812, 0.7637]
Rome Fiumicino	RF	0.8610	[0.8377, 0.8825]	0.7576	[0.7167, 0.7746]
	DT	0.8235	[0.8013, 0.8446]	0.8202	[0.8054, 0.8302]
Naples	RF	0.6875	[0.6446, 0.7110]	0.5662	[0.5369, 0.5971]
	DT	0.6084	[0.5712, 0.6626]	0.4383	[0.3959, 0.4862]
Bologna	RF	0.5811	[0.4084, 0.7353]	0.5906	[0.5067, 0.6690]
	DT	0.5854	[0.3859, 0.7646]	0.5419	[0.4462, 0.6272]

Source: Authors' elaboration on Istat data and FlightAware data

Figure 1. SHAP values computed, for random forest applied to predict passenger arrivals at Rome Fiumicino Airport in June 2023.



In addition, we also compute the SHAP (SHapley Additive exPlanations) values for the RF model, a method introduced by Lundberg and Lee (2017) to explain individual predictions

in machine learning models[7]. Figure 1 illustrates as an example, the SHAP values computed for the random forest regressor, applied to predict passenger arrivals at Rome Fiumicino Airport. Key features influencing the model are highlighted in descending order of importance. For instance, *seats*, *hour*, and *day_of_the_week* exhibit the most significant contributions, as evidenced by their broad spread and concentration of points away from zero. Features such the various airport codes contribute less to the model's predictions, as their SHAP values cluster near zero. This visualization confirms the model's ability to leverage both numerical (e.g., *seats*) and categorical variables (e.g., *day_of_the_week*) to capture the underlying patterns in passenger arrivals effectively.

5. Considerations and future perspectives

ADS-B data provides a valuable and timely supplement to conventional data sources, enhancing the ability to produce comprehensive and real-time air traffic statistics. While it has limitations, particularly in directly capturing passenger and cargo data, these can be addressed through machine-learning estimation algorithms, such as those presented in this study.

To further improve the accuracy and robustness of our passenger and cargo estimation models, several enhancements should be considered. Firstly, integrating a global holiday calendar into the analysis will enable the models to account for fluctuations in passenger and cargo volumes during peak travel periods, such as holidays, when demand typically surges. Additionally, it is important to factor in cancelled flights and codeshare agreements between airlines, as these elements can significantly influence the accuracy of our estimates. Moreover, the number of available seats could be derived from the aircraft type through a web scraping algorithm that extracts this information from sources like Eurocontrol's Aircraft Performance database[8]. Finally, extending the scope of the analysis to encompass all Italian airports and covering all months of the year offer a more comprehensive understanding of flight patterns and demand variations. This broader approach will lead to more reliable predictions across different times and locations, potentially assisting Official Statistics in obtaining real-time data and reducing the costs associated with data collection operations.

However, cargo estimates still require further consideration on how to improve their efficiency. Despite the progress made, enhancing the models' ability to better capture cargo data, especially at airports with lower traffic, remains a crucial area for future work.

References

Breiman, L. (2017). *Classification and Regression Trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, pp. 5-32.

Géron, A. (2019*). Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly.

Karam, R. (2022). *Automatic Detection of Business Data Anomalies with Deep Learning and Application to the ADS-B Protocol* (Doctoral dissertation, Université Bourgogne Franche-Comté).

Lundberg, S. M., Su-In, L. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems.*

Martín Coll, E. (2018). *Flight Plan Information Exchange using SWIM* (Bachelor's thesis, Universitat Politècnica de Catalunya).

[7] In the context of machine learning, SHAP explains a specific prediction by quantifying the contribution of each feature to that prediction. It treats the feature values of a data instance as players in a coalition and evaluates their individual contributions to the model's output.

[8] Available at the https://contentzone.eurocontrol.int/aircraftperformance

doi.org/10.26398/asaproc.00129

A statistical approach to measuring the impact of medication use on the lived experience of patients with inflammatory bowel disease

Alfonso Piscitelli[1], Angela Maria D'Uggento[2], Elena Equatore[2], Maria Beatrice Principi[3]

[1]Department of Agricultural Sciences, Federico II University of Naples, Italy
[2]Department of Economics and Finance, Aldo Moro University of Bari, Italy
[3]Department of Precision and Regenerative Medicine and Ionian Area, Aldo Moro University of Bari, Italy

1. Introduction

Quality of life assessment involves a highly complex assessment of individual well-being within a specific contextual framework. The aim of this study is to investigate the quality of life of people diagnosed with inflammatory bowel disease (IBD) and to assess whether the use of certain medications administered by experimental methods can improve the quality of life of these patients. The participants in this study suffer from IBD, which includes two chronic idiopathic inflammatory diseases: Crohn's disease (CD) and ulcerative colitis (UC).

The primary therapeutic goal for IBD patients is to achieve sustained remission without relapse or the need for surgery while maintaining a high level of well-being. Infliximab (IFX) was the first biologic drug approved by the Food and Drug Administration (FDA) for the treatment of Crohn's disease and ulcerative colitis. In recent years, several studies have supported the commercialization of CT-P13, a biosimilar of IFX SC (subcutaneous formulation). This other formulation has demonstrated equivalent efficacy, bioavailability and safety for maintenance therapy in patients with moderate to severe active Crohn's disease and ulcerative colitis. Subsequently, vedolizumab (VDZ) was also approved, with the subcutaneous formulation (VDZ SC) recently launched. Numerous scientific studies have confirmed that subcutaneous vedolizumab has a high safety and tolerability profile in patients with these diseases. This study focuses on evaluating the impact of novel subcutaneous biologic therapies on quality of life in patients with moderate-to-severe IBD, filling a critical gap in patient-centered outcomes research.

2. Data and methods

All participants in the study completed two specially designed questionnaires to assess their quality of life and their concerns and expectations regarding the switch from intravenous to subcutaneous administration of the drug. The assessment of both quality of life and perception of this switch was carried out at the beginning of treatment (T0) and after six months of therapy (T6). Adverse events following subcutaneous administration of the drug were analysed and divided into two categories: local reactions (such as pain, redness, swelling and induration at the injection site) and systemic reactions (such as fatigue, malaise, headache, muscle pain, joint pain, hives, allergic reactions, diarrhoea, vomiting, anaphylaxis and lymph node swelling).

The McMaster Inflammatory Bowel Disease Questionnaire (IBDQ) was used to assess quality of life, as it is recognized as the most commonly used tool to assess health-related quality of life in patients with inflammatory bowel disease (IBD) (Roseira et al., 2021). The change in patients' quality of life before and after switching to subcutaneous therapy was measured using the validated Italian translation of the IBDQ by Ciccocioppo et al. (2011). To achieve the secondary research objective, a specific questionnaire, the Transition Questionnaire (TQ), was developed by the medical research team to assess patients' concerns and expectations regarding the transition. The TQ consisted of six questions focusing on the patients' psychological state and their potential improvement in coping with daily social activities.

Both the IBDQ and TQ questionnaires were administered to the same group of patients whose therapy was continued to the first interim control point six months after the initial baseline. A six-month period is generally considered appropriate to assess patients' initial perception of changes in quality of life, with subsequent assessments scheduled at six-month intervals from baseline. Patients were divided into two

groups depending on whether they were treated with infliximab (IFX) or vedolizumab (VDZ). At the last time point, patients were also asked whether they had experienced a "general improvement in their well-being" All questionnaires were completed anonymously by the participants who volunteered for the study. The study was conducted between December 2021 and October 2022 in the Department of Gastroenterology of the College of Bari (BA) and in two smaller regional centers, namely the hospitals of Lecce (LE) and Barletta (BAT). Ethical approval was obtained for the study and all patients gave their consent for their data to be used in aggregate form and in full compliance with data protection regulations.

Of the 113 patients who participated in the retrospective study, 108 (95.56%) consented to participate in the survey and successfully completed both the IBDQ and TQ questionnaires at T0. However, only some participants were able to complete the T6 questionnaires as the six-month period of the new treatment had not yet been reached. In addition, patient records with numerous unanswered questions in the T6 questionnaire were removed, which led to the exclusion of these patients from the analysis.

The reliability of the data collected with the IBDQ and TQ questionnaires was assessed using all items both at baseline and at T6. Internal consistency was measured by Cronbach's α. Data were summarized as mean (standard deviation) or median (interquartile range) for continuous variables if they were not normally distributed, while absolute and relative frequencies were calculated for categorical variables.

To control for randomization, the propensity score method (PSM) was used to estimate the causal effects of the two different treatments. This method is particularly useful in contexts where random assignment to an intervention has not been done before (Hong and Raudenbush, 2006; Morgan et al., 2010; Li et al., 2014). In this study, the PSM was used to assess whether patients treated with IFX and VDZ were matched based on similar distributions of relevant covariates. This approach helps to ensure that patients receiving one treatment are not significantly different from those in the other group. The PSM allows researchers to estimate an unbiased treatment effect (Keller and Tipton, 2016). The selection of matching variables was based on their proven relevance to the outcomes of interest, specifically:

– Age and gender: these demographic variables are known to influence both disease progression and treatment outcomes, making them crucial for reducing confounding factors.

– Smoking status: smoking is a well-documented risk factor for inflammatory bowel disease, and inclusion of this factor ensures that differences in smoking prevalence do not confound estimates of treatment effects.

– Disease characteristics: matching of key disease-related variables (particularly severity and duration) accounts for baseline disease heterogeneity that might otherwise bias comparisons between groups.

– Comorbidity: IBD can affect people in different ways and can be associated with numerous physical and psychological comorbidities.

PSM was performed using the Nearest-neighbor algorithm to achieve an optimal balance between the two groups. The threshold for equilibrium was set at <0.1, implying negligible imbalance, based on conventions in the literature (Austin, 2009). This ensures that the covariates are appropriately balanced, which reduces residual confounding. A matching ratio of 1:1 was chosen to maximize the comparability of the groups while maintaining statistical efficiency. A visual inspection of the propensity score distributions was conducted with clinicians to confirm the appropriateness of the matching and the balance achieved for all included covariates.

A total of 46 patients were matched, resulting in 23 pairs. Comparisons were made at both T0 and T6 to assess potential improvements in quality of life and concerns/expectations during the transition to subcutaneous drug administration. As the data were not normally distributed, the Wilcoxon signed-rank test was used to compare the paired data across the two time points. To determine whether differences in treatment affected the improvement in quality of life, the Mann-Whitney U test was used to compare the IFX and VDZ groups.

Predicted propensity score and matching between patients were performed using packages MatchIt (Ho et al., 2011), and matching (Sekhon, 2011) implemented in the statistical computer programming environment R (R Core Team, 2024). Furthermore, the main descriptive statistics as well as the statistical tests between the two treatment groups were performed using the R package.

3. Statistical analysis and results

The demographic and clinical characteristics of all patients who agreed to participate in the study — including local and systemic effects and treatment persistence after six months — are shown in Table 1. The characteristics appear to be fairly balanced in terms of infliximab or vedolizumab treatment, Mayo scores and HBI scores. At baseline, 108 patients completed the IBDQ and TQ questionnaires, while 93 (86.11%) remained available for the second time point.

The majority of participants were male (56%) and the average age was around 42 years. However, patients in the IFX group were significantly younger than those in the VDZ group (p=0.002). The gender distribution was relatively balanced in the IFX group but skewed in the VDZ group. The median disease duration was 12 years and was significantly shorter in the IFX group (p=0.018), which is probably due to the younger age of this group. No significant differences were observed in Mayo scores (p=0.384) or HBI scores (p=0.297) (Table 1). Age and gender were two of the key variables used to match participants who completed the study questionnaires at T6 to the two treatment groups.

Approximately 70% of participants had never smoked and a further 16% were current smokers. Comorbidities were present in 43% of cases, mainly in the IFX group. In terms of the type of IBD, the proportion of patients with ulcerative colitis (UC) was slightly higher in the overall cohort, but three times higher in the VDZ group. Almost 93% of patients reported no local effects, 74% reported no rheumatic effects and an even higher proportion reported no dermatologic or systemic effects.

Table 1. Demographic and clinical characteristics of the patients participating in the study.

	Total sample	*IFX*	*VDZ*	*p-value*
N	108	57	51	
Age (median [IQR])	42 [31 – 57.25]	39 [29 – 49]	53 [34.5 – 64]	0.002
Duration ((median [IQR])	12 [8.75 – 19]	11 [7 – 17]	14 [10 – 21,5]	0.018
Mayo (median [IQR])	1 [1 – 1]	1 [1 – 1]	1 [1 – 1]	0.384
HBI (median [IQR])	0 [0 – 3].	0 [0 – 3]	2 [0 – 3]	0.297
Sex = M (%)	55.56	50.88	60.78	0.851
Smoke = NO (%)	69.44	70.18	72.54	0.691
Comorbidity = NO (%)	42.59	68.42	45.10	0.001
IBD = RCU (%)	60.19	45.61	76.47	<0.001
Chir = NO (%)	84.26	78.95	90.20	0.102
Reuma = NO (%)	75.93	84.21	66.67	0.198
Derma = NO (%)	90.74	91.22	90.20	0.774
Local = NO (%)	92.60	94.49	90.20	1.000
Syst = NO (%)	96.30	96.49	96.08	1.000
Center (%)				
BA	71.30	70.18	72.54	
BT	12.04	10.53	13.73	
LE	16.66	19.29	13.73	

When analysing possible gender differences in IBD, the disease appears to be evenly distributed between men and women (Fig. 1). Smoking appears to be related to the type of disease, as the proportion of smokers is higher in patients with Crohn's disease, while the number of non- smokers predominates in patients with ulcerative colitis. The percentage of smokers was quite low, although most of them were men. There was no significant difference in disease duration between men and women or between smokers (2) and non-smokers (0 = never smoked; 2 = ex-smoker). In terms of age, patients receiving VDZ were generally older (p=0.002), while patients treated with IFX were younger.

This is likely due to the fact that older patients generally had more severe disease and were treated with VDZ, which has been shown to be better tolerated. Examination of the data on disease duration showed that people with a longer history of disease were more likely to receive VDZ and that disease duration was generally longer in older patients. These results were of crucial importance for the comparison of patients in the two treatment groups (IFX vs VDZ). PSM was calculated using the most important covariates, including gender, comorbidities, type of IBD, and disease duration, resulting in 23 pairs of patients (p=0.017). Subsequent comparisons between the two groups (IFX vs VDZ) were performed on a total of 23 pairs of patients who homogeneously compose the two groups which do not differ significantly.

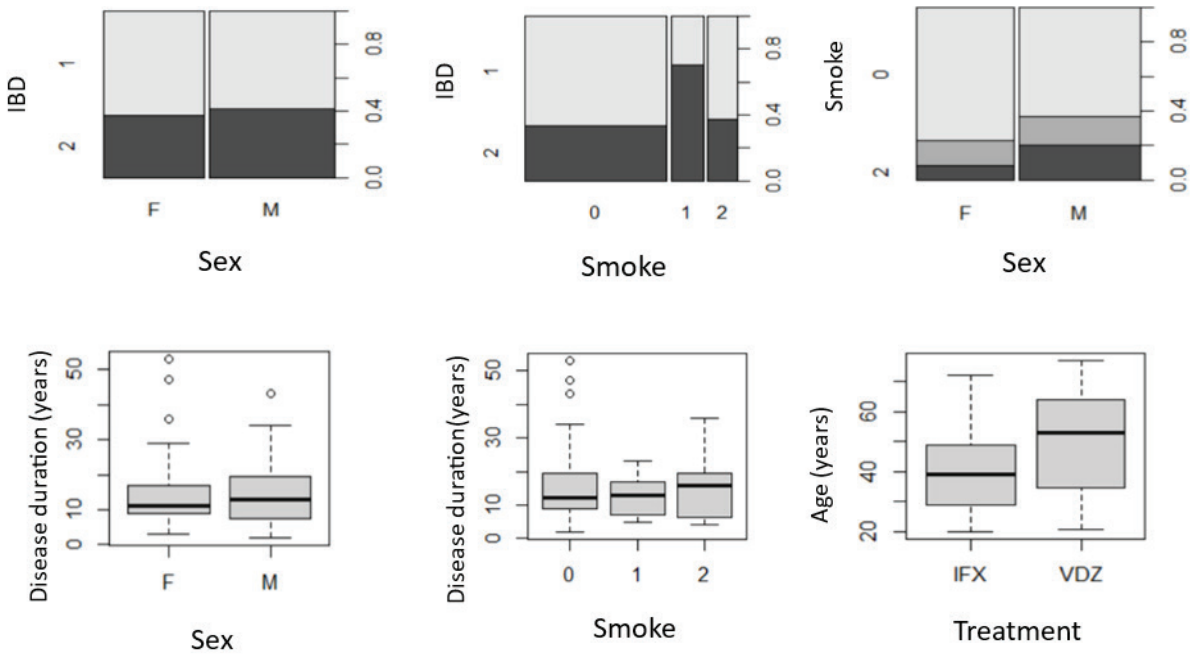


Figure 1. Some relevant characteristics of patients participating in the study.

The demographic and clinical characteristics of the compared patients treated with either infliximab or vedolizumab are shown in Table 2.

Table 2. Demographic and clinical characteristics of the patients interviewed at T6.

	Total	*IFX*	*VDZ*	*p-value*
N	46	23	23	
Age (median [IQR])	41.5 [32 – 56.75]	40 [32 – 50]	53 [33.5 – 65]	0.14
Duration ((median [IQR])	13 [9 – 18.75]	13 [9 – 15]	14 [7.5 – 20]	0.766
Sex= M (%)	52.17	49.13	65.22	0.849
Smoke = NO (%)	69.57	65.22	73.91	0.647
Comorbidity = NO (%)	65.22	86.96	56.52	0.001
IBD = RCU (%)	54.35	21.74	86.96	0.561
Center (%)				
BA	80.43	73.91	86.96	
BT	10.87	17.39	4.34	
LE	8.70	8.70	8.70	

4. Discussion and conclusions

Our hypothesis was that the new method of subcutaneous administration would improve the quality of life of the patients participating in the study, both medically (assessed by the IBDQ) and psychosocially (assessed by the TQ). We then sought to determine whether there were significant differences between the two groups of patients treated with either IFX or VDZ. Table 3 shows the main descriptive statistics and significant differences in relation to the 10 questions of the IBDQ and the 6 questions of the TQ, comparing responses at T0 and T6 for all patients and by treatment group.

Significant positive differences were observed for 5 of the 10 IBDQ questions between T0 and T6 when looking at patients regardless of treatment received (Table 3 - Overall). A comparison of total scores at T0 and T6 also showed that subcutaneous administration generally led to an improvement in quality of life (p=0.029). When comparing IBDQ scores between the two groups, we found that the type of treatment had a significant impact on the improvement in quality of life at six months (p=0.028), while no difference was found at baseline (p=0.956). As shown in the p-values presented in Table 3, patients treated with IFX

reported greater satisfaction than those treated with VDZ.

Table 3. Descriptive statistics for the IBDQ questions at baseline and T6 for all patients and by treatment group (*significant p-values at α=0.05 in bold*)

	Overall			*IFX*			VDZ		
	T0	T6		T0	T6		T0	T6	
	Sum of ratings	Sum of ratings	p-value	Sum of ratings	Sum of ratings	p-value	Sum of ratings	Sum of ratings	p-value
Q1	233	278	**0.000**	119	136	**0.008**	114	142	**0.002**
Q2	282	302	**0.009**	141	148	**0.041**	141	154	0.102
Q3	286	305	**0.010**	139	153	**0.017**	147	152	0.262
Q4	269	270	0.792	131	138	**0.038**	138	132	0.191
Q5	267	281	0.062	132	141	**0.022**	135	140	0.675
Q6	281	287	0.196	140	146	**0.046**	141	143	0.874
Q7	247	261	**0.002**	119	136	**0.016**	128	125	**0.058**
Q8	223	239	**0.019**	109	131	**0.004**	114	108	0.765
Q9	269	248	0.837	137	134	0.366	132	114	0.461
Q10	277	274	**0.054**	141	151	**0.026**	136	123	0.565
Tot	2634	2745	**0.029**	1308	1414	**0.001**	1326	1331	0.631

The same approach was used to compare patients' responses to the TQ questionnaire measuring perceived quality of life at T0 and T6 overall, within the two groups and between the two groups. Table 4 shows that significant positive differences were found for questions Q1 and Q5, without taking into account the possible effects of the treatment received (Table 4 - Overall). In general, patients' quality of life improved by switching to the subcutaneous mode (p=0.001). When comparing the results of the TQ questionnaire between the two groups, we found no significant differences at baseline or after six months (*p-value*=0.991 and *p-value*=0.249, respectively). Analysing the data by treatment, there was an improvement in quality of life in both groups: as shown in Table 4, the total score was significantly higher in either the IFX group (p=0.025) or the VDZ group (p=0.011).

Table 4. Descriptive statistics for the TQ questions at baseline and T6 for all patients and by treatment group (*significant p-values at a=0.05 in bold*)

	Overall			*IFX*			VDZ		
	T0	T6		T0	T6		T0	T6	
	Sum of ratings	Sum of ratings	p-value	Sum of ratings	Sum of ratings	p-value	Sum of ratings	Sum of ratings	p-value
Q1	101	139	**0.011**	51	62	0.303	50	77	**0.017**
Q2	191	197	0.471	97	97	0.861	94	100	0.381
Q3	192	204	0.301	95	100	0.319	97	104	0.617
Q4	205	215	0.070	103	107	0.305	102	108	0.156
Q5	170	217	**0.000**	82	107	**0.011**	88	110	**0.011**
Q6	177	188	0.274	93	95	0.672	84	93	0.304
Tot	1036	1160	**0.001**	521	568	**0.025**	515	592	**0.011**

Although patient convenience alone was not considered a sufficient reason for switching treatment, patient consent remains a mandatory requirement for switching to subcutaneous (SC) administration of medication. There are several reasons why some patients may not prefer subcutaneous administration. These include concerns about receiving less attention from healthcare professionals because they do not visit infusion centers, difficulties with self-administration, and fears that the switch could trigger an exacerbation of their IBD. However, the TRAVELESS program, which facilitated the transition from intravenous vedolizumab to SC vedolizumab, demonstrated high patient satisfaction and provided numerous benefits to the healthcare system. Similarly, rheumatologists have evaluated various therapeutic education programs aimed at improving patient adherence to SC anti-TNFα inhibitors. Strong adherence was found in patients who received individualized therapeutic patient education (TPE).

This study emphasizes that subcutaneous treatment generally improves patients' quality of life. The discrepancy between the results of the IBDQ and the TQ survey can be explained in several ways. First,

unlike the IBDQ, the TQ survey contains questions that require more subjective responses. In addition, the IBDQ responses of the VDZ group were already quite high at T0, indicating that their quality of life was already high. Although there was a general improvement after six months, it was not statistically significant. Therefore, when asked in the survey "Do you think subcutaneous treatment can improve your quality of life?" they gave lower scores as they believed their quality of life was already high and unlikely to improve further.

In contrast, the IFX group, who had a lower quality of life at baseline (reflected in the significant improvement in IBDQ at six months), were more confident at T0 that subcutaneous treatment could improve their quality of life. In addition, as previously demonstrated in the literature, the transition to subcutaneous treatment did not lead to worsening of the disease, and patients generally had no significant difficulties in self-administering the drug.

The results underline the potential of subcutaneous (SC) biologic therapies to improve patients' quality of life and offer more flexibility. Subcutaneous administration reduces the need to visit infusion centers, giving patients more control over their treatment while reducing the burden on the healthcare system. Removing barriers, such as concerns about self-administration, through therapeutic patient education (TPE) programs, such as TRAVELESS, can further improve patient satisfaction and adherence to treatment.

From a policy perspective, switching eligible patients to SC therapies could lead to significant cost savings and more efficient resource allocation. Policy makers should support TPE programs and equitable access to SC treatments to maximize these benefits. Personalized approaches that take into account patient preferences and baseline characteristics are key to optimizing outcomes and aligning with broader health goals.

References

Austin, P. C. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine,* **28**(25), pp. 3083-107.

Ciccocioppo, R. Klersy, C. Russo, M.L., Valli, M., Boccaccio, V., Imbesi V., Ardizzone S., Bianchi Porro, G., Corazza, G.R. (2011). Validation of the Italian translation of the Inflammatory Bowel Disease Questionnaire. *Digestive and Liver Disease,* **43**, pp. 535–541.

Ho, D.E., Imai, K., King, G., Stuart, E.A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software,* **42**(8), 1-28.

Hong, G., Raudenbush, S.W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, **101**, pp. 901–910.

Keller, B., Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, **41**(3), pp. 326–348.

Li, L., Kleinman, K., Gillman, M.W. (2014). A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity. *Journal of Developmental Origins of Health and Disease*, **5**, pp. 435–447.

Morgan, P.L., Frisco, M.L., Farkas, G., Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, **43**, pp. 236–254.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AT (http:/www.R-project.org, accessed 25 June 2024).

Roseira, J., Sousa, H.T., Marreiros, A., Contente, L.F., Magro, F. (2021). Short inflammatory bowel disease questionnaire: Translation and validation to the Portuguese language. *Health and Quality of Life Outcomes*, **19**, pp.1-9.

Sekhon, J.S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, **42**(7), 1- 52.

doi.org/10.26398/asaproc.00130

Artificial intelligence in support of accessibility: The Brain-friend device

Vito Santarcangelo[a], Saverio Gianluca Crisafulli[a], Massimiliano Giacalone[b], Gianfranco Piscopo[d], Emilio Massa[a], Sergio Vitullo[a], Davide Scintu[a], Angelo Romano[a], Girolamo Radosti[c]

[a] iInformatica Srl, Matera
[b] University of Campania 'L. Vanvitelli', Capua
[c] Coralway Srl SB, Matera
[d] University of Naples 'Federico II', Naples

1. Introduction

We live in a context where accessibility and inclusivity are key issues in the socio-technological landscape, ensuring equal access to information and resources is essential to allow each individual to fully participate in society and enjoy their rights. This concept not only concerns the removal of physical barriers, but also of digital and cognitive barriers, where people with visual impairments may have difficulty accessing unstructured online content for reading via screen readers, while those with hearing impairments may encounter obstacles with multimedia content without subtitles or transcripts, just as inclusiveness is not limited to physical disabilities, but also includes cultural, linguistic and cognitive differences.

In this context, generative artificial intelligence offers tools and techniques to improve accessibility and promote inclusivity, thanks to its capacity for machine learning and adaptation to individual needs, providing highly personalised solutions, understanding and interpreting a wide range of data, including text, images and sound, and adapting the way information is presented according to users' preferences and needs (Fernandes et al., 2014).

This paper presents a case study characterized by an accessibility system and a related 'brain-friend' device, an expert system that is able to determine the accessibility profile of a user considering a series of parameters with the aim of generating the content likely requested by the user according to his or her accessibility profile, and also a device for the interaction with the user and the analysis of content according to the accessibility profile determined by the system. Brain-friend represents an important tool for accessibility for digital cultural heritage applications. This paper shows also new possible approaches of the use of "brain-friend" to implement "custobot" experiences and also "human firewall" applications.

2. Brain-friend

In a socio-technological landscape where accessibility and inclusiveness are increasingly important, ensuring that information and resources are accessible to all is crucial. Fair and non-discriminatory access to information is essential for enabling every individual, regardless of their abilities or limitations, to fully participate in society and exercise their rights. Numerous solutions are currently available to enhance accessibility, ranging from traditional assistive technologies to cutting-edge applications of artificial intelligence (AI). AI, in particular, holds the potential to revolutionize accessibility by offering personalized and dynamic solutions tailored to the specific needs of each individual (Bordoni et al., 2013).

Accessibility is not only about removing physical barriers; it also involves eliminating cultural and cognitive obstacles. Everyone should have the right to access information and actively participate in society, regardless of their diversity. For instance, individuals with visual impairments may struggle to access online content that is not properly structured for screen readers, while those

with hearing impairments may find it difficult to engage with multimedia content that lacks subtitles or transcripts. Inclusivity, on the other hand, extends beyond physical disabilities to encompass cultural, linguistic, and cognitive differences.

Artificial intelligence offers a wide array of tools and techniques that can be leveraged to improve accessibility and promote inclusiveness. With the ability to learn and adapt to individual needs, AI-based systems can deliver highly customized solutions that enhance access to information and services. These systems can process and interpret various types of data, including text, images, and sound, and adjust the presentation of information according to the user's preferences and requirements.

The system under study utilizes artificial intelligence in the form of an expert system to determine a user's accessibility profile. The term system here refers to an integrated set of software components, algorithms and hardware that work together to perform specific tasks, using artificial intelligence-based techniques and methodologies. The system considers various parameters collected in a semantic knowledge base, such as age, cultural background, visual, hearing, and cognitive disabilities, as well as other specific individual needs. It also includes a crowd computing module designed to create a collaborative AI model that adapts to the linguistic and terminological profile of the identified user.

Additionally, the system features an AI module that analyzes and interprets textual and multimedia content, along with another AI module that regenerates the analyzed content in a manner that is accessible to the determined user profile. The system can be integrated via API calls into browsers and applications, enabling users to access and comprehend online content according to their specific accessibility needs. Additionally, the system includes an electronic "Brainfriend" device, designed to interact with users and deliver content in alignment with the accessibility profile determined by the system.

This device is equipped with an AI module for interpreting information acquired through a microphone, creating an information model based on what has been learned, and another AI module that provides answers to user queries (e.g., information about a person, a story, or an ongoing encounter). Currently, there are no existing accessibility systems that offer the innovative and personalized solutions required to fully enable every individual to access and participate in the digital world and everyday life.

Within this context, Brainfriend represents a groundbreaking system for accessibility through content interpretation. It features a software module with a user interface that allows for the definition of an accessibility profile (such as age, cultural background, visual impairments, hearing impairments, learning disabilities, or phobias) and a crowd computing module for the creation of a collaborative AI model tailored to the user's terminology and language style. This system regenerates and interprets acquired content, making it accessible according to the user's profile. Integrated into browsers or applications, it allows for the interpretation of web pages, sites, blogs, and news articles, ensuring that content is understandable according to the user's accessibility profile. Moreover, the system can be implemented through an electronic interface (device) that aids in the accessibility of information and decision-making in daily life (Pisoni et al., 2021).

The Brainfriend device assists users in gathering and interpreting information, providing clear answers to their questions (e.g., about a person, a story, or an ongoing encounter). The device guides users in creating their own personalized model, delivering comprehensible responses based on their accessibility profile. In the realm of cultural heritage, Brainfriend may be considered an innovative AI-based system that conveys content to users in a language suited to their accessibility needs (Bordoni and Mele, 2016). By leveraging generative AI, it becomes possible to enhance content enjoyment based on age (child/adult) and cultural background.

Brainfriend is particularly effective in facilitating the understanding of journalistic, scientific, and digital heritage content. For example, each artwork is narrated in an accessible and personalized manner (Abou-Zahra et al., 2018). With Brainfriend, users can also discover information about the artist, historical context, discovery details, and other curiosities through a personalized and

accessible interface. This module is also integrated into virtual reality content consumption scenarios and the "Realverso", the metaverse of sustainability (Chakraborty et al., 2023).

3. Analysis of the potential of Brainfriend

To assess the potential market impact and the response to user needs for Brainfriend, a questionnaire was administered to a convenience sample of 51 users. These participants were distributed across different age groups and were employees of private companies in Italy. All were familiar with generative artificial intelligence and had used it at least once. The survey revealed that 88.2% of respondents perceived a significant level of social marginalization related to cultural levels in their area.

Furthermore, 88.2% of respondents indicated that journalistic information is often incomprehensible to users, regardless of their cultural background, and 94.1% found institutional information similarly difficult to understand. According to the respondents, a "Brainfriend" module capable of interpreting web content according to the user's profile (82.4%), or an electronic device that can listen to people's daily experiences to help them remember events, information, and improve their lives, would be highly useful.

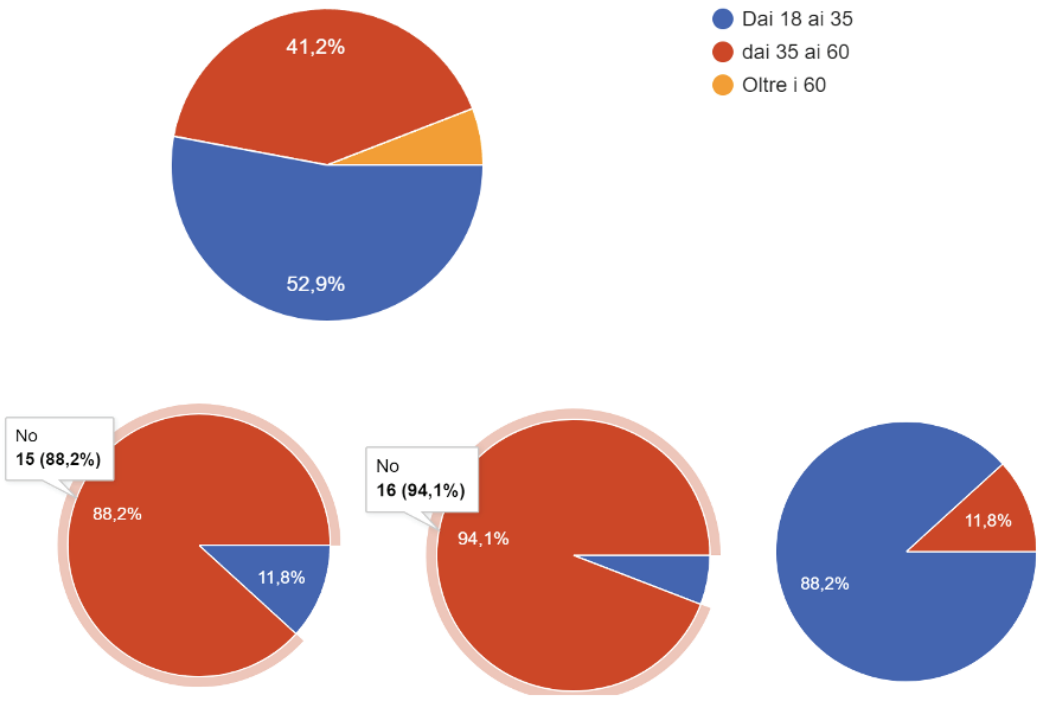


Figure 1 - Results of the field survey

"Brain-friend" operates on a key principle: the customization of content according to the user's needs and level of preparation. We simulated, through different user scenarios, how AI-generated content could be tailored to various target audiences.

Three users are described as examples to highlight how this personalization process might unfold:

Patrick, born in 2012, is a teenager and represents a novice user. The system is designed to provide Patrick with content that is accessible and easy to understand, considering his young age and initial level of competence.

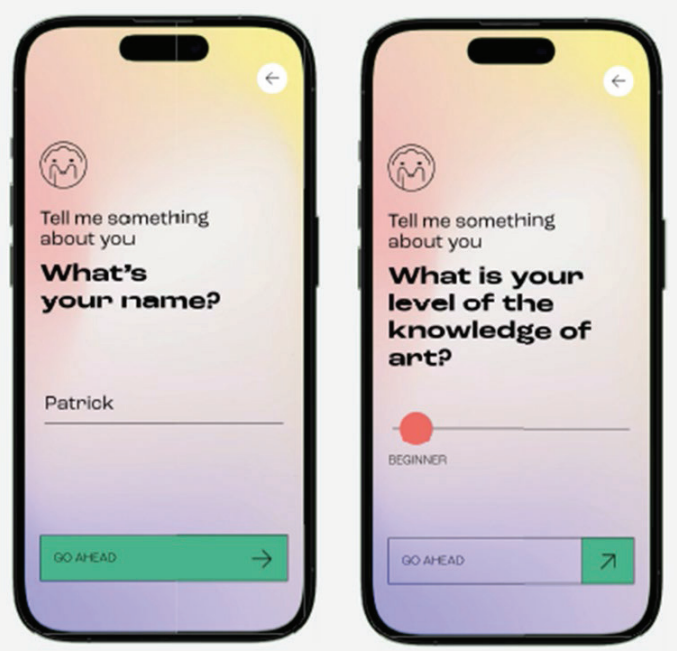


Figure 2 - Patrick

Nicole, born in 1988, is identified as an intermediate user. This suggests that Nicole has some familiarity with the content or system in question but is not yet an expert. For her, the "Brainfriend" system could present content that is slightly more complex than what is provided to Patrick, offering challenges that stimulate her growth and learning without being too difficult to manage.



Figure 3 - Nicole

Olivia, born in 1969, describes herself as an expert. This implies that she has extensive experience and in-depth knowledge in the area of interest covered by the system. For Olivia, "Brainfriend" should generate advanced content that allows her to explore complex topics, catering to her need for a sophisticated and stimulating learning experience.



Figure 4 - Olivia

As highlighted above, "Brainfriend" is an ambitious project aimed at creating a deeply personalized user experience through the use of artificial intelligence, tailoring content not only to the user's level of expertise but also to their specific needs.

4. Application developments and market analysis

Brain-friend represents a crucial tool for improving accessibility, particularly in the context of a new shopping experience. The web of the future must cater not only to users in terms of accessibility but also to AI bots, enabling them to perform actions automatically (e.g., making a purchase).

For example, a user could wish to buy a T-shirt with specific style preferences, size, and budget constraints. An electronic payment tool could then authorize a “custobot” to carry out the search and potentially complete the purchase on the user’s behalf.

To introduce these capabilities, it is essential to rethink web page development, especially in e-commerce, with a focus on AI integration. This approach allows users to delegate tasks, such as searching for and purchasing items, to their "Brainfriend" simply through voice commands.

The challenge now lies in establishing new standards for web development that are sustainable and inclusive for users while also being accessible and interpretable to AI bots (Ara et al., 2022).

Interfacing with tools like Brainfriend would significantly enhance the accessibility of various services for any user group. Additionally, in a world increasingly plagued by online threats, Brainfriend can serve as a key component of the “human firewall,” helping users prevent social engineering attacks that exploit the powerful tools of generative AI.

For instance, it can assess the safety, reliability, and validity of an online offer or sale (e.g., fair pricing, potential health or environmental issues detected through web sources). Thus, AI can be humanity’s first ally in preventing social engineering attacks carried out using advanced digital techniques.

References

Abou-Zahra, S., Brewer, J., Cooper, M. (2018, April). Artificial intelligence (AI) for web accessibility: Is conformance evaluation a way forward?. In *Proceedings of the 15th*

International Web for All Conference, 23-25 April, 2018, Lyon, France, ACM (pp. 1-4).

Ara, J., Cecilia, S.L. (2022). Artificial intelligence in web accessibility: Potentials and possible challenges. *Proceedings of IAC*, 173.

Bordoni, L., Ardissono, L., Barceló, J.A., Chella, A., de Gemmis, M., Gena, C., Sorgente, A. (2013). The contribution of AI to enhance understanding of cultural heritage. *Intelligenza Artificiale*, 7(2), pp. 101-112.

Bordoni, L., Mele, F. (Eds.). (2016). *Artificial Intelligence for Cultural Heritage*. Cambridge Scholars Publishing.

Chakraborty, N., Mishra, Y., Bhattacharya, R., Bhattacharya, B. (2023). Artificial intelligence: The road ahead for the accessibility of persons with disability. *Materials Today: Proceedings*, 80, pp. 3757-3761.

Fernandes, N., Kaklanis, N., Votis, K., Tzovaras, D., Carriço, L. (2014). An analysis of personalized web accessibility. In *Proceedings of the 11th Web for All Conference* (pp. 1-10).

Pisoni, G., Díaz-Rodríguez, N., Gijlers, H., Tonolli, L. (2021). Human-centered artificial intelligence for designing accessible cultural heritage. *Applied Sciences*, 11(2), 870.

doi.org/10.26398/asaproc.00131

Combining textual analysis and IRT to interpret the students' voice

Annalina Sarra, Lara Fontanella, Adelia Evangelista, Tonio Di Battista
Department of Socio-Economic, Managerial, and Statistical Studies, University "G.d'Annunzio" of Chieti-Pescara, Italy

1. Introduction

Over the last decades, numerous studies have investigated student satisfaction at the university level in developed countries. In higher education, student satisfaction is crucial for both the success of institutions and the personal and academic achievements of students (Santini et al., 2017). This is especially important in today's global landscape, where intense competition among institutions and an unprecedented array of choices for students prevail (Paul and Pradhan, 2019). Student satisfaction in higher education is a multifaceted construct, with various dimensions contributing to overall satisfaction and loyalty among students. The literature has identified academic factors as playing a significant role, such as the perceived quality of instruction, the feedback provided by educators, diverse teaching methodologies and the overall quality of learning experiences (Butt and Rehman, 2010; Paul and Pradhan, 2019).

However, the factors contributing to student satisfaction extend beyond the academic realm. Broader institution-wide attributes, including the accessibility of student services, the level of infrastructure support (e.g., transportation, housing, and administrative services), and the quality of facilities, also play a crucial role (Butt and Rehman, 2010; Duque, 2014; Hanssen and Solvoll, 2015). Furthermore, intangible factors such as the institution's reputation and public perception, the degree of student-centeredness, the overall campus atmosphere, and students' personal experiences (Elliott and Healy, 2001) have been recognized as significant contributors to student satisfaction in higher education. Importantly, the quality and nature of students' social connections and relationships with peers, faculty, and staff have also emerged as essential contributors to their satisfaction with the higher education experience (Elliott and Healy, 2001). This aspect has been particularly underscored by the concerns over the loss of social contact and socialization following the suspension of in-person classes due to the COVID-19 outbreak.

Understanding and addressing these diverse dimensions is essential for institutions to enhance student satisfaction and, ultimately, their overall educational experience and success. Surveys have been a useful tool for universities worldwide to identify these needs and expectations, enabling them to develop policies and procedures in response to these findings.

In this vein, this study examines the experiences of over 5,000 students enrolled at an Italian university. The students were asked to share their views on three key areas: their degree program, their overall university experience, and the city where they reside. They rated their responses on a five-point scale. Additionally, each section of the survey featured two open-ended questions, encouraging students to identify positive aspects and suggest areas for improvement in each dimension. To assess the students' satisfaction, we integrated text mining and item response theory (IRT) methodologies. The remainder of the paper is organized as follows: Section 2 gives insights about participants, instruments and methodology for data analysis. Section 3 presents the main results, while the final Section offers the main conclusions from the analysis.

2. Materials and methods

2.1 Participants

Students who participated in this research were enrolled in one of the 68 bachelor's, master's, or

single--cycle degree programs offered by the participating institution. These students were invited to participate in an online survey in the middle of the 2023/2024 academic year. The response rate for the questionnaire was approximately 25%, with 5,280 out of 20,814 enrolled students participating in the survey, indicating that one out of four students took part. The variability across degree programs stands at 8.3% (interquartile range). The response rate is higher for master's degree programs (28.5%) compared to bachelor's degree programs (23.6%). High rates (around 30%) are recorded for some single-cycle degree programs. In this sample, 3994 were female, with the most represented age group being between 18 and 22 years old (49%). The majority of respondents, 39.1%, are enrolled in courses within the social sciences field. The health sciences area represents the second most common choice, with 28.4%. Another significant percentage of respondents, 19.4%, are enrolled in courses within the humanities field. The scientific field has a smaller representation, with 13.0% of respondents.

2.2. Instruments

Drawing from existing literature, the data collection involved an anonymous questionnaire divided into four sections. The initial section was aimed to collect socio-demographic data, encompassing gender, age, high school education, city and region of residence, enrolled degree program, and factors influencing university choice. A key focus of the questionnaire was to investigate students' satisfaction regarding educational offerings, university experience, and aspects of city life. Subsequently, the following three sections delved into various factors, such as teacher preparedness, practical learning opportunities, skills development for real-world challenges and career objectives, interaction with peers, use of social media for academic discussions, utilization of university resources, sense of belonging, friendships, and the role of university life in socialization. Furthermore, the questionnaire assessed aspects concerning the size, cost of living, suitability for study, availability of services, job opportunities, cultural events, and leisure activities of the campus city. Each aspect was represented by a single item in the student satisfaction measure, with students responding on a 5-point rating scale. In addition to the quantitative measures, each of the key sections of the survey featured two open-ended questions. These qualitative prompts were designed to encourage students to share their personal perspectives and experiences. The first open-ended question in each section asked students to identify and describe the positive aspects within that particular dimension while the second open-ended question in each section then invited students to suggest areas for improvement. This gave participants the opportunity to voice their ideas and recommendations for enhancing various aspects of their university life and the local environment.

2.3. IRT modeling on the itemized questionnaire

Responses to the itemized questionnaire were used as input for statistical analysis through Item Response Theory (IRT; Lord (1980)). IRT is a powerful modeling approach employed to evaluate the psychometric properties of survey questionnaires that utilize categorical (ordered or unordered) response formats. One of the most relevant IRT models for analyzing ordinal responses, such as Likert-type scales, is the graded response model (GRM; Samejima (1969)). To assess how well each scoring option or response category reflects the underlying latent trait, student satisfaction in our case, we adopted the multidimensional extension of Samejima's GRM. Consider a k −item instrument composed of m subscales, where each subscale contains k_v polytomous response items measuring a single latent dimension. The probability that person i selects category c for item j in the v-th subscale ($v = 1, 2, \dots, m$) is given by:

$$\mathrm{P}\left(Y_{vij} = c \mid \theta_{vi}, \alpha_{vj}, \delta_j \right) = \Phi\left(\alpha_{vj}\theta_{vi} - \delta_{j,c-1}\right) - \Phi\left(\alpha_{vj}\theta_{vi} - \delta_{j,c}\right), \qquad (1)$$

where Y_{vij} represents the response of person i to item j in subscale v, and θ_{vi} is the latent trait level of person i in that subscale. The parameter α_{vj}, known as the discrimination

parameter, indicates how strongly item j within subscale v is associated with the latent trait. Higher values of α_{vj} suggest that the item is more effective in differentiating between respondents with varying levels of the trait, while lower values indicate a weaker association. Unlike simpler models that assume equal discrimination across items, the GRM allows for each item to have a unique discrimination value, making it more flexible in accommodating variations in item sensitivity. The parameter δ_j represents the threshold parameters for item j, which denote the points along the trait continuum where respondents have a 50% likelihood of selecting a particular response category or higher. These thresholds position each item along the latent trait, reflecting the "difficulty" or "severity" of each response level. Finally, Φ denotes the cumulative distribution function of the standard normal distribution, applied to model the probability structure of the responses. To estimate both item and person parameters, we used Bayesian estimation with MCMC simulation techniques. This method proved effective for parameter estimation and model comparison. For each item, we obtained posterior distributions for the discrimination and threshold parameters, and calculated latent satisfaction scores for each respondent. These scores represent the predicted position of respondents on different satisfaction traits, helping us analyze satisfaction patterns across various aspects of the university experience.

2.4. Text mining of open-ended questions

To automatically analyze the extensive text generated from open-ended questions, we created a textual corpus and applied text mining and topic modeling techniques. Specifically, we utilized latent Dirichlet allocation (LDA; Blei et al. (2003)) to uncover the underlying topics within the documents. LDA is a well-known unsupervised, generative probabilistic method that views each document as a mixture of random latent topics. Each document is composed of "bag-of-word" topics, where each topic has a discrete probability distribution indicating the likelihood of word occurrences. An important feature of the LDA algorithm is its document-to-topic assignment process. During this phase, each word is allocated to a topic, and each document is associated with one or more topics. Formally, LDA can be described as follows. Assume we have a corpus consisting of D documents, where each document d is a sequence of N words, $w_d = (w_{d,1}, w_{d,2}, \dots, w_{d,N})$, and each word is an item from a vocabulary indexed by $\{1, \dots, V\}$. Additionally, there are K latent topics, $\beta_{1:K}$, defined as distributions over the vocabulary. The generative process of LDA can be described by the following joint distribution of both hidden and observed variables:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})$$

The topic proportions for document d-th are θ_d, with $\theta_{d,k}$ being the proportion for topic k in document d. The topic assignments are z_d, where $z_{d,n}$ is the topic assigned for the n-th word in document d. Both the topic proportions and the topic distributions over the vocabulary follow a Dirichlet distribution. As the posterior distribution, $p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D})$, is intractable, approximate inference methods or variational algorithms are commonly applied.

We refined our analysis by examining linguistic variations across different levels of participant satisfaction. Students were grouped into four categories based on quartiles of latent scores from a GRM model, reflecting satisfaction across the two dimensions analyzed. Through keyness analysis, we identified terms strongly associated with each of the four sub-corpora. This statistical method compares observed and expected frequencies, highlighting terms that are distinctive to a specific group. The significance of the discrepancy between observed and expected terms is assessed using the chi-square χ^2 metric. The analysis revealed the most

characteristic terms for each satisfaction level, offering insights into the language patterns and themes in teachers' responses across the quartiles.

3. Results

This section presents the findings from the multidimensional GRM and the analysis of open-ended responses. The GRM results were obtained by implementing a custom MATLAB routine on the first two dimensions of the questionnaire: *Degree program* and *Overall university experience*. Table 1 shows the posterior means and 95% credible intervals (CIs) for the discrimination parameters of items in these dimensions. All estimates are statistically significant, as the CIs do not include 0. In the *degree program* dimension, the aspects with the largest impact on satisfaction were *skills provided* (posterior mean = 1.866) and *professional goals* (posterior mean = 1.173). For the *overall university experience* dimension, *student community* showed the highest influence (posterior mean = 1.198), followed closely by *socialization opportunities*. The latent trait correlation between the two dimensions was 0.499, indicating a moderate relationship between satisfaction with the degree program and the broader university experience. Item-person maps (not shown here) revealed higher satisfaction levels across both dimensions. Key aspects of satisfaction include effective professional goal achievement and faculty preparation in the degree program, alongside meaningful friendships and diverse socialization opportunities within the university experience.

Table 1: Posterior means and 95% credible intervals for discrimination parameters of items in the Degree program and Overall university experience dimensions.

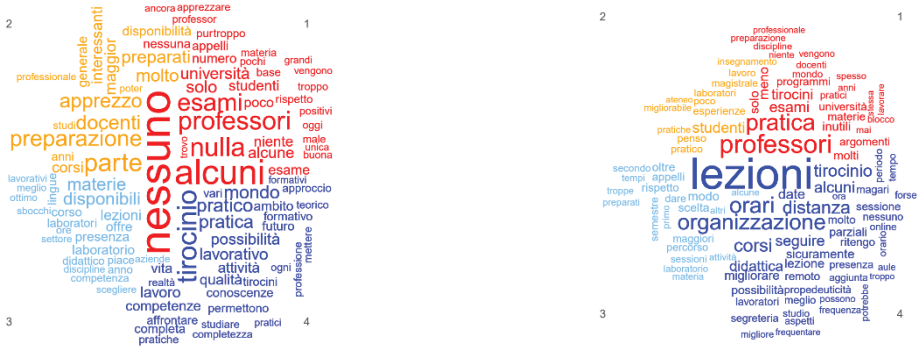
Item	Posterior Mean	95% CI Lower	95% CI Upper
Satisfaction with the degree program			
Well-prepared professors	0.659	0.614	0.698
Practical learning	1.088	1.035	1.151
Skills provided	1.866	1.680	1.913
Professional goals	1.173	1.073	1.192
Satisfaction with university experience			
Studying with others	0.665	0.612	0.694
Social media	0.528	0.461	0.540
University resources	0.643	0.595	0.670
Student community	1.198	1.083	1.196
Meaningful friendships	0.897	0.852	0.951
Socialization	1.143	1.085	1.223
Student involvement	0.958	0.917	1.026
Latent trait correlation	0.499	0.472	0.528

To gain deeper insights, we analyzed responses to open-ended questions using the quanteda package (Benoit et al., 2018) in the R environment. Students were divided into four groups based on quartiles of satisfaction levels, derived from latent scores in the GRM analysis. The keyness analysis highlighted terms strongly associated with each satisfaction group (see Figure 1 and Figure 2). The comparison word cloud in the left panel of Figure 1 visually emphasizes the linguistic differences between satisfaction groups regarding the positive aspects of the degree program. In the lowest quartile, terms such as “none,” “nothing,” and “nothingness” suggest disappointment, while “internship”, “competence”, and “future” emerged in the highest quartile, reflecting appreciation for professional preparation and

opportunities. Additionally, using the LDA technique, we identified topics highlighting strengths and areas for improvement in both dimensions. Results of the topic analysis for the degree program are shown in Fiure 3.

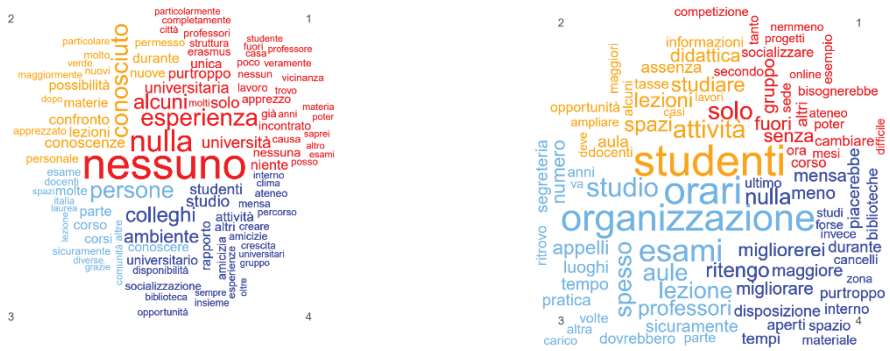
4. Concluding remarks

The combination of itemized questionnaire and open-ended questions in our study offers valuable insights to help the university improve student satisfaction. Our research highlights key strengths of the *degree program*, such as strong practical opportunities like internships and lab work, which are essential for career preparation. The diverse course offerings provide a well-rounded education, blending theoretical knowledge with practical insights. Faculty dedication and expertise also contribute to a positive student experience. However, improvements are needed in addressing heavy workloads, increasing flexibility in course selection, and securing more field-specific internships. Streamlining the organization of practical activities, expanding exam scheduling options, and ensuring accessible materials for remote learners are crucial steps to enhance the program and better prepare students for their careers. Regarding the *university experience*, strengths include accessible faculty, a dynamic learning environment, meaningful peer connections, and a well-structured campus that supports both academic and social life. Areas for improvement include optimizing educational organization, simplifying administrative procedures, boosting student engagement in activities, expanding extracurricular options, enhancing support for first-year students, increasing study space availability, and upgrading services such as dining and bars to meet student needs.



(a) Positive aspects **(b)** Areas of improvement

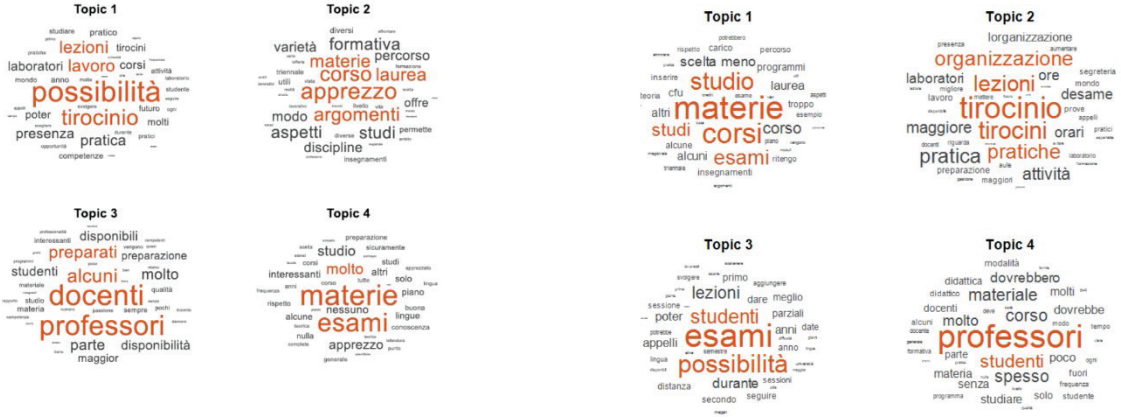
Figure 1: Word cloud representing respondents' opinions on the degree program based on their level of satisfaction



(a) Positive aspects **(b)** Areas of improvement

Figure 2: Word cloud representing respondents' opinions on the degree program based on their level of satisfaction

Figure 2: Word cloud representing respondents' opinions on university experience based on their level of satisfaction



(a) Positive aspects **(b)** Areas of improvement

Figure 3: The topics of evaluations on the positive aspects and areas for improvement of the Degree Program

References

Benoit, K., , K., Wang H., Nulty, P., Obeng, A., Müller, S., Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, **3** (30):774.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* **3**(1), pp. 993–1022.

Butt, B., Rehman, K. (2010). A study examining the students satisfaction in higher education. *Procedia - Social and Behavioral Sciences*, **2**(2), pp. 5446–5450.

Duque, L. (2014). A framework for analysing higher education performance: Students' satisfaction, perceived learning outcomes, and dropout intentions. *Total Quality Management & Business Excellence*, **25**(1-2), pp. 1–21.

Elliott, K. M., Healy, M. A. (2001). Key factors influencing student satisfaction related to recruitment and retention. *Journal of Marketing for Higher Education*, **10**(4), pp. 1–11.

Hanssen, T., Solvoll, G. (2015). The importance of university facilities for student satisfaction at a Norwegian university. *Facilities*, **33**(13-14), pp. 744–759.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Paul, R., Pradhan, S. (2019). Achieving student satisfaction and student loyalty in higher education: A focus on service value dimensions. *Services Marketing Quarterly*, **40**(3), pp. 245–268.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, **34**(1), pp. 1–97.

Santini, F., Ladeira, W., Sampaio C., da Silva Costa, G. (2017). Student satisfaction in higher education: A meta-analytic study. *Journal of Marketing for Higher Education*, **27**(1), pp. 1–18.

doi.org/10.26398/asaproc.00132

Exploring teachers' professional identity: A case study utilizing structural equation modeling and text mining in a multimodal approach

Annalina Sarra[a], Maila Pentucci[b], Lara Fontanella[a]
[a] Department of Socio-Economic, Managerial, and Statistical Studies, University "G.d'Annunzio" of Chieti-Pescara, Italy
[b] Department of Humanities, Arts, and Social Sciences, University "G.d'Annunzio" of Chieti-Pescara, Italy

1. Introduction

The profile of the professional teacher has long been a focus of research that underscores its dynamic and multidimensional nature (Gracia et al., 2021). Numerous researchers have sought to delineate teacher professionalism by developing frameworks that outline essential and indispensable competencies. These frameworks not only help define the process of professionalization but also aid in structuring both initial and ongoing training programs. Notably, influential models include Perrenoud's ten competencies for teaching (Perrenoud, 2002), Altet's synthesis of necessary practical and declarative knowledge (Altet, 2006), and Le Boterf's fields of construction (Le Boterf, 2010). Their role spans a technical dimension linked to practical wisdom, enabling them to address unprecedented and urgent situations in action, and a cultural dimension (Robichaud et al., 2015) that imbues their work with human, moral, and political significance (Pentucci and Cioci, 2023).

Professional identity is thus a dynamic element, necessitating not only the continual realignment of specific competences to evolving contexts but also operational autonomy and situational intelligence within a framework of ongoing restructuring (Pentucci, 2018). Certainly, factors both internal and external to practice influence the self-perception of the teacher's identity. Reflexivity on one's actions, a primary characteristic of being a teacher, should make individuals aware of the essential dimensions of practice outlined by Vinatier (2013): the pragmatic issue, linked to teaching skills; the relational issue, related to pedagogical and interpersonal skills; and the epistemic issue, associated with knowledge of the discipline and reference knowledge.

However, professional identity is also shaped by the environment in which these practices are exercised. It is important to remember that this identity is both individual and collective, necessitating ongoing mediation between a teacher's personal choices and conceptions and those of the reference practice community. This involves a dialogue between "genre", i.e., common references accumulated over time, and "style", i.e., individual peculiarities (Clot and Faïta, 2000).

Furthermore, the teacher's role as a social actor, who performs non-routine intellectual operations and assumes personal responsibility, must not be overlooked, especially given the ethical, political, and value conflicts that beset the school (Rushton and Reiss, 2021). Starting from this framework, it was intriguing to analyze teachers'self-perceptions of their individual identities by focusing on these two categories of factors: internal factors, noting the importance they attribute to their didactic, pedagogical, and disciplinary expertise exercised in the classroom; and external factors, linked to life and work contexts, the relevance of the taught discipline to their self-image, and their accumulated experience. The aim is to understand how the role, function, and concrete actions of teaching are directly experienced and interpreted by those involved, amidst the complexities of contemporary education.

The rest of the paper is structured as follows: Section 2 provides details on participants, instruments, and the methodology used for data analysis. Section 3 presents the main findings, while

the concluding section provides a discussion of the results and summarizes the conclusions drawn from the analysis.

2. Materials and methods

2.1 Instruments

In this study, we utilized a revised version of the questionnaire developed by Beijaard et al. (2000) which is intended to investigate various dimensions of teachers' professional identity. The questionnaire consists of two main sections. The first section includes general questions about the background variables of the teachers: gender, age, years of experience as a teacher, and school level (early childhood education level, primary, middle and high school level). This section aims to gather demographic and professional background information that can help contextualize the responses.

The second section has 18 items measuring teachers' professional identity and 16 items evaluating contextual, experiential, and biographical influences, all on a four-point scale. This provides a comprehensive assessment of the factors shaping teachers' professional identities.

Additionally, three open-ended questions were included to enrich the collected data. These questions aimed to encourage teachers to reflect on their learning regarding the disciplinary, educational, and pedagogical aspects of their role over their careers. Specifically, the questions prompted teachers to consider: (1) what they have learned about their subject matter expertise, (2) their pedagogical strategies, and (3) the broader educational contexts in which they work.

2.2. Participants

For our study we utilized a convenience sampling approach, distributing an online questionnaire to teachers from educational institutions within the Marche and Abruzzo regions. This sampling method allowed us to efficiently gather data from a readily available pool of participants. A total of 509 questionnaires were collected, yielding a comprehensive dataset for analysis. The respondents were predominantly female, comprising 79% of the total sample (403 females; 106 males). This gender distribution highlights the significant representation of women in the teaching profession within this context.

In terms of professional engagement, the majority of participants (44%) worked at the high school level, followed by 23% at the middle school level, 23% at the primary school level, and 11% at the early childhood education level. The mean number of years teachers had been in service was 12. This average indicates a relatively experienced cohort, which may have implications for their professional development needs and their perspectives on teaching practices and policies. Furthermore, 38% of the respondents were newly hired, reflecting a significant proportion of early-career teachers in the sample. Additionally, the average age of the teachers was 46 years. This age distribution suggests a diverse mix of early, mid, and late-career teachers, which can provide a broad range of insights into the profession.

2.3. Structural equation model

Causal relationships and associations among the latent variables in the administered questionnaire were explored using structural equation modeling (SEM; Bollen (1989)), a widely-used quantitative technique in fields such as education, psychology, sociology, and business. The widespread adoption of SEM can be attributed to its versatility and ability to explore and test complex relationships between multiple variables, including both observed and unobserved (latent) constructs. By integrating concepts from path analysis, factor analysis, and multiple regression, SEM allows researchers to assess the extent to which their hypothesized models are supported by empirical data, facilitating theory development and validation. One of the key distinguishing features of SEM is its ability to encompass both measurement and structural models. The measurement model in SEM establishes a connection between the observed variables and the underlying latent variables, allowing researchers to account for

measurement error and understand how the observed indicators are related to the unobservable, latent constructs. The structural model examines the relationships between the endogenous variables (which can be either latent or observed) and the exogenous variables (also either latent or observed), enabling researchers to test theoretical hypotheses about the directional influences and reciprocal relationships among the variables of interest.

What sets SEM apart from other analytical techniques is its ability to examine a large number of relationships simultaneously while accounting for measurement error. In our model specification, we assume the presence of three endogenous latent variables (specifically, "*didactic*", "*pedagogy*", and "*subject*") and three exogenous latent variables ("*context*", "*experience*" and "*biography*"). The observed variables for teacher i ($i = 1,...,N$) are represented by X_{il}, $l = 1,...,L$, which measure the exogenous latent variables (θ_q), ($q = 1,...,Q$); while the observed variables that measure the endogenous latent variables η_q are denoted by Y_{ik}, $k = 1,...,K$.

The measurement model, where the underlying variables are linked to the observed indicators through a threshold model, can be expressed as:

$$Z_{il}^{(x)} = \lambda_l^{(x)}\eta_i + \varepsilon_i^{(x)}, \text{ with } X_{il} = c \text{ if } \tau_{l,c-1}^{(x)} \leq Z_{il}^{(x)} \leq \tau_{l,c}^{(x)} \quad (1)$$

$$Z_{ik}^{(y)} = \lambda_k^{(y)}\theta_i + \varepsilon_i^{(y)}, \text{ with } Y_{ik} = c \text{ if } \tau_{k,c-1}^{(y)} \leq Z_{ik}^{(y)} \leq \tau_{k,c}^{(y)} \quad (2)$$

In the above expressions, we denoted with $Z_{il}^{(x)}$ and $Z_{ik}^{(y)}$ the underlying continuous variables, whereas $\lambda_l^{(x)}$ and $\lambda_k^{(y)}$ are the discrimination parameters, with $\tau_{l,c-1}^{(x)}$ and $\tau_{k,c-1}^{(y)}$ being the threshold parameters and $\varepsilon_i^{(x)}$ and $\varepsilon_i^{(y)}$ the normally distributed errors. Moving on to the structural model, this can be written as:

$$\boldsymbol{\eta}_i = \boldsymbol{\beta}'\boldsymbol{\theta}_i + \boldsymbol{g}_{wi} + \boldsymbol{u}_i \quad (3)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{g}$ are vectors of regression coefficients, w_i is the vector of observed variables, and u_i is a normally distributed error term.

2.4. Textual analysis of open-ended survey responses

To explore textual data arising from open-ended survey questions, we utilized text mining techniques. Initially, to ensure that the textual analysis is accurate and efficient, minimizing noise and maximizing the relevance of the extracted information, we followed a series of pre-processing steps. Next, to gain insights into how thematic content varies among different groups of teachers, we linked textual analysis with SEM-derived factors. Specifically, we segmented teachers into four groups based on quartiles of factor scores obtained from SEM for the three endogenous variables (didactic, pedagogy, and subject). This segmentation allowed us to observe how significant terms identified by keyness analysis vary across these groups. It is worth noting that keyness analysis relies on a relative frequency comparison of n-gram usage between two groups: a reference group and a target or study group. The keyness scores for an n-gram are determined using the χ^2 test statistic, which tests the null hypothesis that the n-gram is used with the same frequency in both groups. Essentially, the keyness score indicates how distinctive a term is to the reference group. Common words used equally by both groups will not have a high keyness score; by contrast, words with high keyness scores are used relatively more frequently by the target group than the reference group.

The framework for analyzing textual data is completed by applying latent Dirichlet allocation (LDA; Blei et al. (2003)). This technique aims to identify the underlying semantic dimensions, often referred to as "topics" or "concepts," present in the text. LDA assumes that each document is a mixture of a small number of topics, with each topic being a probability distribution over words in the vocabulary. An important assumption of the LDA model is the "exchangeability" or "bag-of-words hypothesis," which means that the importance of a word is

related to its frequency rather than its order within documents. The more often two words co-occur in a document, the more likely they belong to the same topic (Aggarwal and Zhai, 2012).

3. Results

The SEM model was estimated utilizing the R *lavaan* library, as developed by Rosseel (2012). Regression results are summarized in Table 1. It was observed that the latent exogenous variable "context" consistently exerts a strong positive influence across all three dimensions ("pedagogy", "didactic", and "subject"), emphasizing its critical role in teaching practices. Additionally, the variable "age" emerged as a significant factor, positively impacting all three dimensions, indicating that more experienced teachers may exhibit superior pedagogical, didactic, and subject-specific practices. Also, the "school level of teacher" positively influenced both the pedagogy and didactic dimensions. In particular, it is emerged the formative role of early educational levels in shaping effective teaching approaches.

The findings also reveal specific impacts of gender and biography on the examined teaching dimensions. As previously pointed out, keyness analysis and LDA provide a robust framework for exploring textual data and uncovering meaningful patterns that contribute to a richer understanding of the underlying content and its variations among distinct groups. For instance, the keyness analysis shows how the "pedagogy" dimension of the questionnaire influences teachers' priorities regarding the pedagogical aspects of their work, with a focus that ranges from student learning and growth (Group 4) to classroom adaptation and management (Group 1) (Figure 1-a). The topic analysis conducted on the question "*Regarding the pedagogical dimension of my teaching role, during my teaching career, I have mainly learned that...*" identified two main topics: "Creating a positive and effective learning environment" and "Adapting teaching methods to meet student needs", as can been inferred from Figure 1-b.



Figure1: Textual analysis for open-ended questions related to Pedagogy dimension

4. Discussion and concluding remarks

To identify significant influencing factors, we decomposed professional identity into core dimensions within didactic practice: the pedagogical and relational, didactic and pragmatic, epistemological, and mastery of subject matter. Beijaard et al. (2000) identifies three main subjective influencing factors. First is the school context, which includes both the classroom environment and broader school culture. This context encompasses the shared conceptions, norms, and values that foster a unique approach to teaching. Our study found that context significantly shapes the pedagogical aspect of professional identity, with the school community sharing pedagogical values and approaches that become a collective heritage. This supports concept of a community of practice, suggesting that professional identity is shaped through interactions with colleagues and institutions.

Table 1: Results of SEM regression

	Estimate	Std.Err	z-value	P(>|z|)
Pedagogy			-	
Biography	-0.090	0.101	0.889	0.374
Context	0.652	0.092	7.096	**0.000**
Experience	0.157	0.108	1.448	0.148
Male	-0.124	0.080	-1.541	0.123
Age	0.012	0.006	2.166	**0.030**
Years of service	0.005	0.005	0.975	0.330
Kindergarten teacher	-0.052	0.122	-0.431	0.666
Primary school teacher	0.207	0.094	2.191	**0.028**
Middle school teacher	0.128	0.084	1.528	0.126
Newly hired teacher	-0.030	0.085	-0.352	0.725
Didactic				
Biografy	-0.181	0.098	-1.843	0.065
Context	0.617	0.090	6.855	**0.000**
Experience	0.125	0.113	1.106	0.269
Male	-0.223	0.084	-2.653	**0.008**
Age	0.017	0.006	3.041	**0.002**
Years of service	0.007	0.005	1.371	0.170
Kindergarten teacher	0.033	0.110	0.296	0.767
Primary school teacher	0.191	0.091	2.097	**0.036**
Middle school teacher	0.155	0.086	1.791	0.073
Newly hired teacher	0.055	0.082	0.671	0.502
Subject				
Biography	-0.115	0.059	-1.971	**0.049**
Context	0.401	0.071	5.612	**0.000**
Experience	0.128	0.069	1.868	0.062
Male	0.001	0.047	0.019	0.985
Age	0.009	0.003	2.849	**0.004**
Years of service	0.002	0.003	0.705	0.481
Kindergarten teacher	0.004	0.074	0.056	0.955
Primary school teacher	-0.015	0.049	-0.300	0.764
Middle school teacher	-0.055	0.048	-1.142	0.253
Newly hired teacher	0.049	0.048	1.026	0.305

Other influencing factors include personal biography, which affects identity related to subject mastery. Experience and educational background significantly influence a teacher's passion and commitment, contributing to self-perception and motivation. Field experience is crucial for consolidating subject mastery, though teachers often see it as less relevant to pedagogical aspects. This reflects a belief in the foundational role of professional history, while the complexities of modern educational contexts prevent reliance solely on experience. The rise of unexpected challenges has normalized emergency situations in classrooms. Age, as an objective factor, also influences teachers' self-perception of identity, contributing more to self-awareness and overall identity than to perceptions of specific competencies.

Textual analysis of open-ended responses, using topic modeling and keyness analysis, revealed themes that deepen our understanding of teachers' professionalism. In particular, themes related to pedagogical and relational aspects emerged, underscoring the importance of creating positive learning environments and adapting methodologies to meet students' needs. Examining the results of keyness and topic analysis across all three aspects of Vinatier's framework (Vinatier, 2013) proves valuable. This approach not only aids in profiling teaching staff but also supports the reorganization of professional development paths to better align with teachers' needs and beliefs.

References

Aggarwal, C.C., Zhai, C. (2012). *Mining Text Data (1st ed.)*. Springer, New York.

Altet, M. (2006). Le competenze dell'insegnante-professionista: saperi, schemi d'azione, adattamenti ed analisi. In L. Paquay, M. Altet, E. Charlier, P. Perrenoud, eds, *Formare gli insegnanti professionisti. Quali strategie?*, PUR Rennes, pp. 31-44.

Beijaard, D., Verloop, N., Vermunt, J.D (2000). Teachers' perceptions of professional identity: An exploratory study from a personal knowledge perspective. *Teaching and Teacher Education.*, **16**, pp. 749–764.

Blei, D. M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3(1),** pp. 993–1022.

Bollen, K.A. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, Oxford, England.

Clot, Y., Faìta, D. (2000). Genres et styles en analyse du travail: Concepts et méthodes. *Travailler*, **4**, pp. 7–42.

Gracia, E.P., Rodrıguez, R.S., Pedrajas, A.P., Carpio, A.J. (2021). Teachers' professional identity: Validation of an assessment instrument for preservice teachers. *Heliyon*, **7**(9), pp. 115-123.

Le Boterf, G. (2010). *Professionnaliser. Construire des ParcoursPpersonnels de Professionnalisation*. Eyrolles, Paris.

Pentucci, M., Cioci, G. (2023). Expectations and beliefs about professional identity: A study on teachers enrolled in TFA. *QTimes, Journal of Education, Technology and Social Studies*, **15**(3), pp. 294–311.

Pentucci, M. (2018). *I formati pedagogici nelle pratiche degli insegnanti.* FrancoAngeli, Milano.

Perrenoud, P. (2002). *Dieci nuove competenze per insegnare*. Anicia, Roma.

Robichaud, A., Tardif, M., Perlaza, A.M., LeVasseur, L., Jeffrey, D. (2015). *Sciences sociales et théories critiques dans la formation des enseignants*. Presses de la Université de Laval, Quebec.

Rosseel, Y. (2012). An r package for structural equation modeling. *Journal of Statistical Software*, **48**(2), pp. 1–36.

Rushton, E.A., Reiss, M.J. (2021). Middle and high school science teacher identity considered through the lens of the social identity approach: A systematic review of the literature. *Studies in Science Education*, **57**(2), pp. 141-203.

Vinatier I. (2013). *Le travail de l'enseignant*. DeBoeck, Bruxelles.

doi.org/10.26398/asaproc.00133

Selecting sample households with high quality addresses for population census: Problems for field survey network, but bias in the estimates?

Emanuela Scavalli
Directorate for Social Statistics and Population Census, National Institute of Statistics, Rome, Italy

1. Introduction

The new permanent population and housing census combines administrative sources with annual sample surveys in all Italian municipalities (Bernardini et al., 2022). This approach improves the accuracy of the data, while verifying and complementing the information in the registers (Agafiţei et al., 2015).

One of these surveys, the List survey, annually samples one million households in 2,530 municipalities (Gallo and Zindato, 2021). Sample selection prioritizes high-quality addresses to facilitate field operations, but this choice raises questions about potential sampling bias.

This paper presents an analysis of the impact of this choice on the estimation process and assesses whether it actually leads to a bias in the selected sample.

2. The List survey

The List survey design is based on annual samples of one million households collected for the period 2022-2026 in all municipalities, which are divided into two groups:

i) the AR Municipalities that participate yearly (1.188 units);

ii) the NAR Municipalities (6.716 units), appropriately stratified, that participate once over five years (1.343 sample units yearly).

The survey involves a sample of households to which a letter is sent notifying them of being drawn from the sample and an internet address where they can connect to fill out the questionnaire. Subsequently, when the survey starts, reminder operations are planned on the field by surveyors who show up at the addresses of the not yet respondent households asking them to complete the questionnaire.

The selection of the households is random from the Basic Population Register (RBI), but certain constraints are imposed to facilitate the field operations, for logistical problems. One of these constraints is to select households from those with high quality addresses, in order to have a greater guarantee that the households will receive letters and that the surveyors will be to find the households for reminders. The addresses present in the population register are obtained by integrating different registers (Solari et al., 2023). The addresses are categorised as follows:

- High quality (A): complete and correct
- Medium quality (B): partially incomplete or ambiguous
- Low quality (C): incomplete or incorrect (e.g. street and house number missing), often unusable without intervention.

In order to avoid losses due to address inaccuracy, it was decided to randomly select the sample from households with A-quality addresses. On the one hand, this implies a sample selection that a priori excludes a part of the households that may have different characteristics from those with high quality addresses, thus biasing the estimates.

Unfortunately, in several municipalities, especially small ones, the percentage of households with high quality addresses is not high, so households with low quality addresses (B) were extracted, assuming the risk of falling. However, it was found that the response of these households was very

similar to that of the sample with high quality addresses, indicating that even if the address was not complete, the letters were likely to arrive and the households to respond.

3. Quality addresses in population register

In order to assess possible bias among the census respondents, we examined whether the proportion of households with low quality addresses had different individual characteristics from those with high quality addresses, or whether they might be specific to certain areas.

Table 1 shows the distribution of individuals in the 2,531 sampled municipalities in 2022, corresponding to 41,331,748 individuals. It can be seen that the percentage of individuals living in high quality addresses is 97.4%.

Table 1. Population in 2022 sampled municipalities by address quality

Address Quality	Population	
	Absolute Values	Percentages
A	40.261.606	97,4%
B	873.584	2,1%
C	196.558	0,5%
Total	**41.331.748**	**100%**

Comparing the distributions of some characteristics (sex, citizenship and age group) on all individuals and on individuals who have high quality addresses, no significant differences are generally observed.

Table 2 Population by sex and high quality address (HQ)

Sex	Total Population	Population in HQ	% Total Population	% Population in HQ
Female	21.288.194	20.763.855	51,5%	51,6%
Male	20.043.554	19.497.751	48,5%	48,4%
Total	**41.331.748**	**40.261.606**	**100,0%**	**100,0%**

Table 3. Population by citizenship and high address quality (HQ)

Citizenship	Total Population	Population in HQ	% Total Population	% Population in HQ
Italian	37.453.573	36.457.209	90,6%	90,6%
Non Italian	3.878.175	3.804.397	9,4%	9,4%
Total	**41.331.748**	**40.261.606**	**100,0%**	**100,0%**

Table 4. Population by age and high quality address (HQ)

Age	Total Population	Population in HQ	% Total Population	% Population in HQ
<15	4.836.045	4.694.024	11,7%	11,7%
15-34	8.492.120	8.249.857	20,6%	20,5%
35-54	11.541.864	11.224.064	27,9%	27,9%
55-74	11.090.274	10.814.769	26,8%	26,9%
>=75	5.371.445	5.278.892	13,0%	13,1%
Total	**41.331.748**	**40.261.606**	**100,0%**	**100,0%**

4. Municipalities with low quality address in the 2022 sample

In order to assess the potential for bias in the estimates at the municipal level, the distributions for each sampled municipality have been analysed. It was observed that the percentage of high-quality addresses exhibits considerable variability, which may also influence the relative distributions by sex, age, and citizenship. Table 5 presents the distributions of the 20 municipalities with a percentage of high-quality addresses below 50%.

Table 5. Population by address quality in municipalities with a low-quality address

Municipality	Address Quality (AV)				Address Quality (%)		
	A	B	C	tot	A	B	C
Valtournenche	0	2.229	0	2.229	0%	100%	0%
Funes/Villnöß	425	12	2.100	2.537	17%	0%	83%
Ussita	176	163	38	377	47%	43%	10%
Poggiodomo	32	67	0	99	32%	68%	0%
Caposele	1.139	2.143	17	3.299	35%	65%	1%
Mirabella Eclano	447	6.817	1	7.265	6%	94%	0%
Morra De Sanctis	52	1.090	0	1.142	5%	95%	0%
Senerchia	39	689	23	751	5%	92%	3%
Contursi Terme	1.354	1.920	0	3.274	41%	59%	0%
Teggiano	65	7.244	136	7.445	1%	97%	2%
Santo Stefano di Sessanio	40	69	6	115	35%	60%	5%
Cortino	257	343	1	601	43%	57%	0%
Muro Lucano	1.973	2.918	90	4.981	40%	59%	2%
Brancaleone	442	2.928	18	3.388	13%	86%	1%
Ferruzzano	263	493	20	776	34%	64%	3%
Lipari	5.632	6.442	651	12.725	44%	51%	5%
Nissoria	1.413	1.421	14	2.848	50%	50%	0%
San Mauro Marchesato	291	1.608	3	1.902	15%	85%	0%
Francavilla Angitola	619	1.170	60	1.849	33%	63%	3%
Spilinga	678	673	83	1.434	47%	47%	6%

As can be seen in the table, there are some municipalities with few high quality addresses, suggesting that the excluded part of the population is clearly significant.

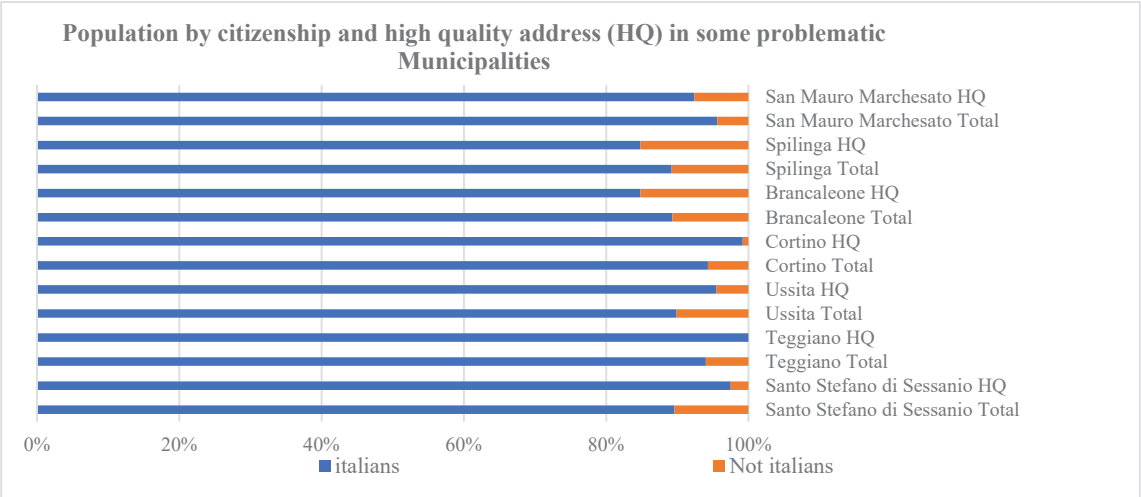
Looking at the trend in the distribution by sex of the total population and the population with high quality addresses in these municipalities (Table 6), we see that there is actually quite a lot of variability between the two distributions.

Table 6. Population by sex and high quality address (HQ)

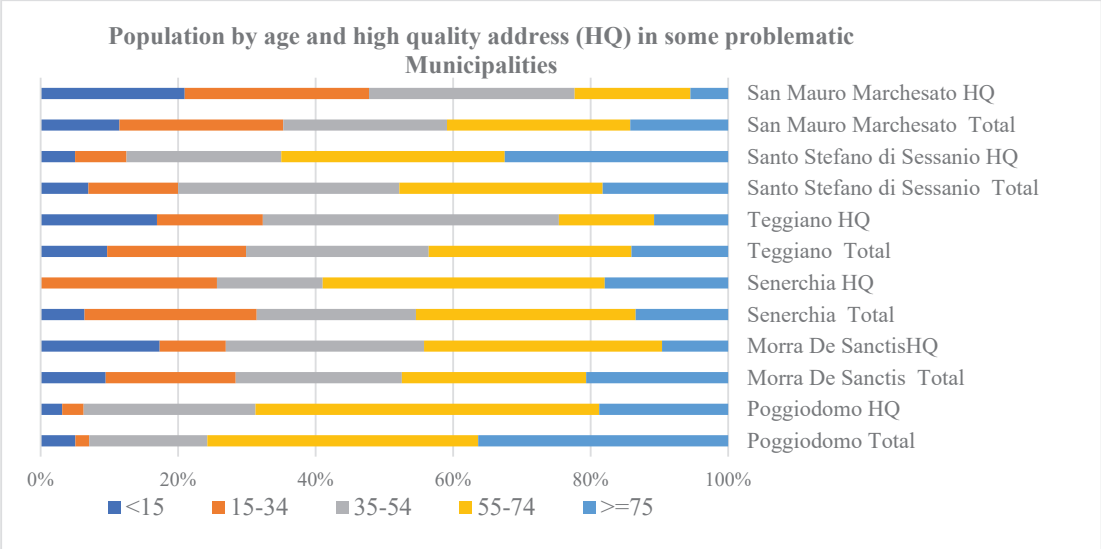
Municipality[1]	Population		Population in HQ		%Population		%Population in HQ	
	F	M	F	M	F	M	F	M
Funes/Villnöß	1.266	1.270	223	202	49,9%	50,1%	52,5%	47,5%
Ussita	193	184	81	95	51,2%	48,8%	46,0%	54,0%
Poggiodomo	44	55	13	19	44,4%	55,6%	40,6%	59,4%
Caposele	1.691	1.608	589	550	51,3%	48,7%	51,7%	48,3%
Mirabella Eclano	3.691	3.574	243	204	50,8%	49,2%	54,4%	45,6%
Morra De Sanctis	590	552	20	32	51,7%	48,3%	38,5%	61,5%
Senerchia	345	406	21	18	45,9%	54,1%	53,8%	46,2%
Contursi Terme	1.650	1.624	707	647	50,4%	49,6%	52,2%	47,8%
Teggiano	3.852	3.593	32	33	51,7%	48,3%	49,2%	50,8%
Santo Stefano di Sessanio	61	54	22	18	53,0%	47,0%	55,0%	45,0%
Cortino	276	325	118	139	45,9%	54,1%	45,9%	54,1%
Muro Lucano	2.509	2.472	1.020	953	50,4%	49,6%	51,7%	48,3%
Brancaleone	1.725	1.663	217	225	50,9%	49,1%	49,1%	50,9%
Ferruzzano	414	362	136	127	53,4%	46,6%	51,7%	48,3%
Lipari	6.276	6.449	2.848	2784	49,3%	50,7%	50,6%	49,4%
Nissoria	1.419	1.429	699	714	49,8%	50,2%	49,5%	50,5%
San Mauro Marchesato	982	920	148	143	51,6%	48,4%	50,9%	49,1%
Francavilla Angitola	920	929	305	314	49,8%	50,2%	49,3%	50,7%
Spilinga	717	717	334	344	50,0%	50,0%	49,3%	50,7%

[1] Excluding Valtournenche, which has no high quality addresses

Even looking at the distribution by citizenship, we note that the population in high quality addresses presents a very different distribution of Italians and foreigners, especially in the municipalities of Santo Stefano di Sessanio, Teggiano, Ussita and Cortino, where the foreigners are less represented, while in Brancaleone, Spilinga and San Mauro Marchesato the foreigners are over represented. To simplify, we do not report all the data, but only those with particularly irregular distribution in the following graphic.



Comparing the distributions by age group, the same distortions are observed in the municipalities reported above, especially in the municipality of Senerchia which has no individual among the high quality addresses in the 0-14 age class.



5. Large municipalities with low address quality in the 2022 sample

Focusing on the 21 large municipalities with more than 50,000 inhabitants and a percentage of low quality addresses of more than 3%, a significant presence of low quality addresses can be observed in Messina, Corigliano-Rossano, Marano di Napoli and Caltanissetta. In these cases, it was more difficult to place the addresses in the registers, given the quality of the registers from which

they were taken. Note that in Reggio di Calabria there is a high percentage of addresses of very poor quality C.

Table 7. Population by quality address in some large Municipalities

Municipality	Address quality (AV)				Address quality (%)		
	A	B	C	tot	%A	%B	%C
Messina	153.739	59.343	11.591	224.673	68%	26%	5%
Corigliano-Rossano	55.151	19.730	2.064	76.945	72%	26%	3%
Marano di Napoli	41.016	7.120	4.558	52.694	78%	14%	9%
Caltanissetta	51.518	7.866	280	59.664	86%	13%	0%
L'Aquila	49.705	6.558	583	56.846	87%	12%	1%
Reggio di Calabria	130.186	4.209	13.021	147.416	88%	3%	9%
Vittoria	56.431	4.220	2.152	62.803	90%	7%	3%
Giugliano in Campania	61.509	4.935	462	66.906	92%	7%	1%
Ragusa	67.195	5.152	410	72.757	92%	7%	1%
Foggia	135.205	12.223	23	147.451	92%	8%	0%
Latina	119.100	8.667	38	127.805	93%	7%	0%
Bari	293.404	11.261	6.852	311.517	94%	4%	2%
Mazara del Vallo	47.708	3.085	142	50.935	94%	6%	0%
Manfredonia	51.652	2.258	269	54.179	95%	4%	0%
Bitonto	50.798	1.984	592	53.374	95%	4%	1%
Firenze	344.089	18.809	40	362.938	95%	5%	0%
Matera	56.416	3.137	67	59.620	95%	5%	0%
Gela	69.773	2.476	449	72.698	96%	3%	1%
Teramo	119.807	4.427	970	125.204	96%	4%	1%
Benevento	56.498	2.265	208	58.971	96%	4%	0%
Taranto	182.203	7.978	261	190.442	96%	4%	0%

The distributions by sex do not present particular distortions (due to space limitations, these are not shown here), as well as the distribution by citizenship, for which it can be noted only a decrease in the incidence of foreigners in Vittoria, Manfredonia, Foggia and Ragusa, while in Messina and Corigliano-Rossano the percentage of foreigners in high quality addresses is greater than the total

Table 8. Population by address quality and citizenship in some large municipalities

Municipality	Vittoria		Manfredonia		Messina		Corigliano-Rossano		Foggia		Ragusa	
	Total	HQ	Total	HQ	Total	HQ	Total	HQ	Total	HQ	Total	HQ
Italians	88,5%	90,7%	96,9%	98,3%	94,4%	93,1%	89,5%	88,6%	93,9%	94,8%	91,8%	92,5%
Not italians	11,5%	9,3%	3,1%	1,7%	5,6%	6,9%	10,5%	11,4%	6,1%	5,2%	8,2%	7,5%

6. Total non-responses in 2022 sample

Looking at the distribution of the respondents in the 2022 sample (Table 9) by quality address, it can be seen that there are no particular non-response rates in the sample with respect to B low quality addresses; which means that households probably received the questionnaire and were able to complete it, even if the addresses were not formally complete. The non-response is likely to be due to other reasons not related to the quality of the address. The situation is different for quality C addresses, which often have almost no addresses at all, but even among these (very few cases) a percentage of respondents was found.

It should be stressed that in small municipalities, even if the addresses are not complete, it is probably easier to identify the households, both because they are very small areas and the postman is likely to know the households living there personally.

Table 9. Respondents and non-respondents in 2022 sample by address quality

Quality Address	Respondents	Non respondents	Total sample	% Non respondents
A	1.947.103	272.591	2.219.694	12%
B	4.515	218	4.733	5%
C	177	61	238	26%
Total	**1.951.795**	**272.870**	**2.224.665**	**12%**

7. Conclusions

The overall percentage of high quality addresses is particularly favourable.

The response rates for low quality addresses are comparable to those for high quality addresses, indicating that letters are still being delivered and that the behaviour of the two populations is not significantly different.

The number of municipalities with a high concentration of low-quality addresses is relatively small, comprising approximately 20 out of 2,530. Despite the presence of these municipalities, the gender, age and nationality distributions in these areas do not appear to have a significant impact on the estimates. This is also due to the fact that a constrained weighting estimator (Deville & Särndal, 1992) was used in the sample weighting stages: this estimator constrains the distributions of these three variables to the total distribution resulting from the register, effectively correcting any potential bias in the estimates.

In summary, high quality addresses dominate the sample distribution, ensuring effective data collection with minimal bias. Corrective weighting techniques further address potential biases and support reliable population estimates.

Future strategies should focus on improving register quality, particularly in municipalities with lower quality address concentrations, to increase coverage and operational efficiency.

References

Agafiţei, M., Gras, F., Kloek, W., Reis, F., Vâju, S. (2015). Measuring output quality for multisource statistics in official statistics: Some directions. *Statistical Journal of the IAOS* 31, pp. 203-211.

Bernardini, A., Chieppa, A., Cibella, N., Gallo, G., Solari, F., Zindato, D. (2022). Evolution of the Italian permanent population census. Lessons learnt from the first cycle and the design of the permanent census beyond 2021, Economic Commission for Europe, *Conference of European Statisticians, Group of Experts on Population and Housing Censuses*, Geneva, Switzerland, 21-23 September 2022.

Deville, J.C., Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, (418), pp. 376-382 .

Gallo, G., Zindato, D. (2021). Italy: The combined use of survey and register data for the Italian permanent population census count in UNECE, *Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses* .https://unece.org/statistics/publications/CensusAdminQuality.

Solari, F., Bernardini, A., Cibella, N. (2023). Statistical framework for fully register based population counts, *METRON*, 10.1007/s40300-023-00244-5.

Zhang, L. C. (2022). Complementarities of survey and population registers, in N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri J., L. Teugels (eds.), Wiley StatsRef: Statistics Reference. https://doi.org/10.1002/9781118445112.stat08352

doi.org/10.26398/asaproc.00134

Statistical analysis of smartphone mobility data for air quality assessment

Muhammad Haroon Shaukat [a], Francesco Finazzi [b]
[a] Department of Engineering and Applied Sciences, University of Bergamo, Italy.
[b] Department of Economic Sciences, University of Bergamo, Italy.

1. Introduction

Population growth and urbanization have led to widespread pollution from industry, waste, heating, and traffic, causing cardiovascular and respiratory diseases (Yurtseven et al., 2018). To address this, personal exposure must be assessed by combining environmental data with individual mobility tracking. High-resolution location data improves exposure accuracy, linking pollution to health impacts more effectively (Finazzi and Paci, 2019).

Air quality is particularly crucial in cities like Istanbul, which has the highest population and busiest traffic in Turkey. Over the last five decades, the population of Istanbul has expanded sixfold, transforming it into a major metropolis both regionally and globally. Ministry of Environment and Urban Development of Turkey highlighted that gas emissions in the city exceed recommended levels. The primary contributors to this pollution are vehicular traffic and industrial activities. Additionally, Istanbul's location, which connects continents, results in heavy maritime traffic, further exacerbating emission levels (Bucak et al., 2021). Interestingly, restrictions on human mobility during the COVID-19 pandemic not only helped reduce the virus's spread but also improved air quality by curbing industrial production, transportation, and traffic (Zhu et al., 2020). This study examines the link between human mobility and air quality, focusing on the mediating role of individual movement patterns.

2. Data

This study analyses particulate matter with a diameter of 10 micrometers PM_{10} data obtained from an open-access platform[1], and the dataset covers measurements from thirteen air quality monitoring stations across Istanbul, including locations such as Aksaray, Besiktas, Esenyurt, Kadikoy, Kartal, Esenler, Sariyer, Sultangazi, Sile, Silivri, Umraniye, Uskudar, and Basaksehir. The data spans from March 1, 2023, to April 30, 2023, consisting of daily observations, recorded in micrograms per cubic meter. Secondly, the ERA5-Land dataset, with a 9x9 km latitude-longitude grid, was used due to its accurate representation of historical climate conditions. This study uses four climate variables covering the area around Istanbul's air quality monitoring stations, aligned with the air quality data, obtained from the Copernicus Climate Change Service[2]. Additionally, the individual smartphone mobility dataset of 443,786 users is considered, covering the same time period as previously mentioned, and was obtained from the Earthquake Network Project (www.sismo.app). This project operates a global earthquake early warning system using networks of smartphones (Finazzi, 2016).

3. Methodology

3.1 Fixed rank kriging

[1] https://aqicn.org/historical/#city:turkey/istanbul
[2] https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

Fixed rank kriging (FRK) is a scalable spatio-temporal modeling and prediction framework that deals with large datasets and can easily handle data with varying spatial supports. FRK relies on a spatial random effects (SRE) model, where a spatially correlated mean-zero random process is decomposed into a linear combination of spatial basis functions with random coefficients, along with an additional term to capture the fine-scale variation of the random process. Let $\{Y(s): s \in D\}$ denote the spatial process, where s indexes represent the location within the spatial domain D. According to Zammit-Mangion and Cressie (2017), the classical spatial statistical model is as follows,

$$Y(s) = t(s)^T \alpha + \upsilon(s) + \xi(s), \quad s \in D \tag{1}$$

where s belongs to the spatial domain, $t(s)$ represents the spatially referenced covariates, α is associated regression coefficients with covariates, $\upsilon(s)$ indicates the small-scale, spatially correlated random effect, and $\xi(s)$ is a fine-scale almost spatially uncorrelated random effect. The latent process' structure can be described in terms of a linear combination of a fixed number of spatial basis functions which constitutes the SRE model as follows:

$$\lambda(s) = \sum_{l=1}^{r} \phi_l(s)\eta_l + \xi(s), \quad s \in D \tag{2}$$

where $\eta = (\eta_1, ..., \eta_r)$ is an r-variate random vector, and $\phi(.) = (\phi_1(.), ..., \phi_r(.))$ indicating the r-dimensional vector of pre-specified spatial basis functions (SBF). Further, the comprehensive information on spatial domain discretization into basic areal units (BAUs), parameter estimation, and the observational equation can be found in Zammit-Mangion and Cressie, (2017).

3.2 Performance measurement

A common approach for validating the accuracy of spatial-temporal models is to use the leave one out cross validation (LOOCV) approach. This method involves sequentially omitting data from one location (typically a station) at a time and estimating the value at that omitted location using the remaining data. According to Robinson and Metternicht (2006), the root mean square error (RMSE) is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M} [Y(s_i, t_i) - \hat{Y}(s_i, t_i)]^2}{M}}, \tag{3}$$

where $(M; i = 1, ..., M)$ indicating the total number of observations, $Y(s_i, t_i)$ are the actual space-time points, and $\hat{Y}(s_i, t_i)$ are the predicted space-time points.

4. Results and discussion

In the FRK setting, the first step is to create the BAUs N which can be either equally sized grids or hexagons. BAUs partition the spatial domain into discrete regions, allowing the fine-scale variation of latent processes to be modelled effectively. This approach allows to one use internal covariates or external. Therefore, the external covariates are assumed to be associated with each BAU, and the fine-scale variability can be incorporated in the form of weighted

considering the variability of spatial latent processes. As far as the weights need to be equal to the number of BAUs.

The dimensions of the BAUs are considered to be 7x7 km, and therefore 124 BAUs are created per day. Consequently, over 61 days, a total of 7564 BAUs were obtained. The next step involves integrating ERA5-Land variables, available at a 9x9 km resolution, with the BAUs at a 7x7 km resolution, and this integration was achieved using a K-Nearest Neighbors (KNN) interpolation approach combined with inverse distance weighting (IDW). Additionally, BAUs are temporally replicated for each unique day to enable temporal interpolation. With 124 BAUs per day over 61 days, this results in a total of 7564 BAUs. For each day, ERA5-Land variables are interpolated for each BAU using IDW, where closer neighbors exert more influence on the interpolated values. This method ensures that both spatial and temporal correlations are effectively captured in the interpolation process. To include individual human mobility patterns as a covariate in the FRK model, a threshold of 3.5 km was applied. Since the BAUs have a uniform resolution of 7x7 km, any time a smartphone user passed within a 3.5 km radius of a BAU, it was considered that the user was present or had passed through that area on a specific date. This presence was marked with a "1" (indicating presence), while all other locations where the user neither passed through nor stayed within the 3.5 km radius of BAU were marked with a "0" (indicating absence). If a smartphone user passed through or stayed within the specified 3.5 km radius at the same location multiple times in a day, their presence was recorded cumulatively for that location and date. This process was repeated for each date to track the presence and absence of smartphone users across the spatial area and this covariate is reflected in the study as the presence of smartphone users.

It is important to describe that this study uses the three different types of basis functions such as bisquare, exponential, and Matern32, and for different combinations of basis function models, the centers of the temporal basis functions (TBFs) are set regularly using a sufficiently large-scale parameter (σ = 2) days and resulting 30 TBFs. For spatial basis functions (SBFs), two resolutions were considered with fixed scale parameters: a radius of 25 km for the first resolution and a radius of 15 km for the second resolution. SBFs are irregularly distributed across the entire geographical area of Istanbul, Turkey, covering the entire spatial domain of interest. As a result, 37 SBFs were created, with 4 corresponding to the first resolution and 33 to the second resolution. The use of two levels of spatial resolution in the basis functions can enhance the capability to model spatial correlations. Generally, a smaller resolution can capture local correlations, while a larger resolution can model more general correlations.

Afterward, both basis functions were combined for the spatial-temporal setting. This was achieved using the tensor product, resulting in 1110 combined spatial-temporal basis functions r obtained. Now, there are 5551 observations represented by footprints m with each footprint associated with a specific measurement error σ_{ε}^{2}. Nevertheless, the uncertainty for each measurement is assumed to be uniform across all footprints, set at a value of one, ensuring consistency throughout the spatial domain. Further, the expected maximization (EM) algorithm is employed with 30 iterations with a tolerance of 0.1 to estimate parameters. These settings are generally applied to fit all seven FRK models, each involving different combinations of basis functions. To ensure the validity of all seven models, the study employs the LOOCV criteria and then root mean square error (RMSE) was computed between the actual and predicted observations for each station and also computed average RMSE for all models. Further detail is described in Table 1.

From Table 1, it can be observed that the average RMSE shows small differences across all models, but the FRK model with Bisquare spatial and temporal basis function demonstrated the highest accuracy as compared to other combinations of models. Ultimately, this model was

chosen to be applied across the entire dataset for predicting PM_{10} concentrations at unobserved locations in Istanbul, Turkey. The remaining steps of the model development followed the previously described procedure. The estimated intercept value is $30.51\,\mu g/m^3$. Since the covariates are externally linked to the BAUs, the FRK model does not provide parameter estimates for the covariates.

Table 1: Comparison of FRK models by LOOCV w.r.t to spatial-temporal basis function

Model	M1	M2	M3	M4	M5	M6
Spatial and Temporal Basis Function	Bisquare Spatial and Bisquare Temporal Basis Functions	Bisquare Spatial and Exponential Temporal Basis Function	Bisquare Spatial and Matern32 Temporal Basis Function	Exponential Spatial and Matern32 Temporal Basis Function	Matern32 Spatial and Matern32 Temporal Basis Function	Matern32 Spatial and Exponential Temporal Basis Function
LOOCV	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
Aksaray	12.1065	12.2447	11.9755	12.3789	12.3344	12.5441
Umraniye	9.3920	9.7646	9.7105	9.0167	9.2403	9.3756
Basiktas	13.2308	13.3406	13.3271	14.8686	14.7408	14.2934
Esenler	10.5712	10.2615	10.5730	9.7162	9.7065	9.7316
Esenyurt	18.4579	18.7876	18.9342	20.5982	20.0287	19.4824
Kartal	10.7937	10.9611	10.7347	10.4567	10.4169	10.8187
Kodikoy	13.2275	13.0958	13.3579	11.8482	11.9578	12.2006
Saryeri	18.4394	19.1571	19.9308	18.9089	19.6090	18.7359
Sile	17.0695	15.9573	15.5948	15.3947	15.3134	15.6900
Silivri	11.4963	12.0427	13.1218	9.9174	11.3160	10.9611
Sirinevler	8.4078	8.4574	9.0209	8.4999	8.6039	7.8630
Sultangazi	17.0238	17.2307	17.6058	18.9261	18.9885	18.4150
Basakeshir	17.8019	18.2404	18.5482	18.5216	19.0364	18.5397
Average	13.6937	13.8109	14.0335	13.7732	13.9456	13.7424

Nine random days were chosen from between two-month study period, as illustrated in Figure 1. Generally, the PM_{10} concentrations are stable, but on March 20, 2023, several areas experienced significantly higher levels. These included regions north of Aksaray and Esenler, as well as Basaksehir, Sultangazi, and Sariyer. Similarly, elevated levels were observed near the Umraniye and Kartal air quality monitoring stations in the southeast. The noticeable concentrations were also recorded between Sile and Esenyurt. On March 25, 2023, similar concentration patterns were observed in the same areas mentioned earlier, but the levels were lower than those recorded on March 20, 2023.

On the remaining days, there were no noticeable fluctuations in PM_{10} concentration, though small increases were observed in the previously mentioned areas. A clear connection between the presence of smartphone users and PM_{10} concentration was not observed, even when applying the same model fitting procedure without considering the covariate as the

presence of smartphone users. There are several possible reasons for this. First, the two-month study period may be too short for spatial-temporal modeling, as it limits the capture of long-term trends and seasonal patterns. Second, this study relied on smartphone data from 443,786 users, which includes a low sampling rate as coordinates are recorded at irregular timestamps, and approximately 0.44 million is a small sample size relative to the broader population of Istanbul. To investigate the general relationship, the correlation between the total presence of smartphone users and PM_{10} concentration is computed as shown in Figure 2.

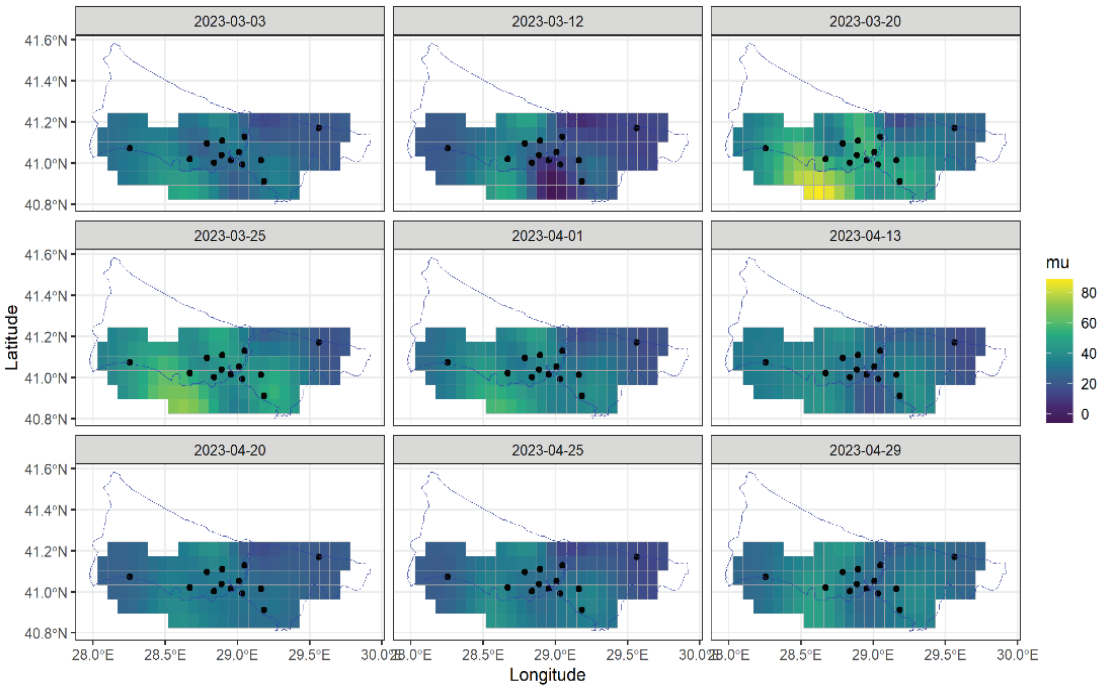


Figure 1: Prediction of average PM_{10} for randomly selected days

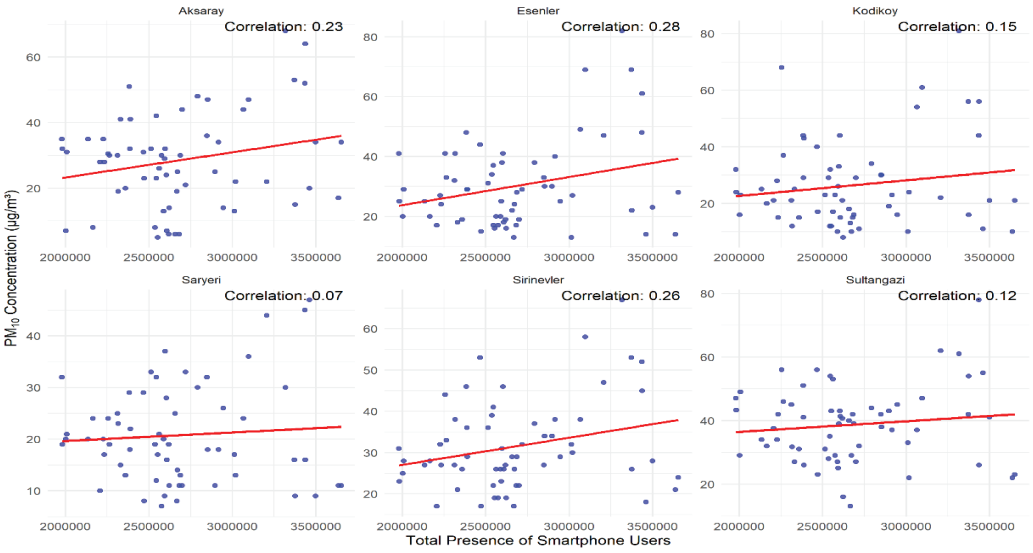


Figure 2: Scatter plot of total presence of smartphone users and PM_{10} concentration

From Figure 2, all air quality monitoring stations are located in the central area of Istanbul. It is observed that there is an overall very low correlation between the total presence of smartphone users and PM_{10} concentrations of all air quality stations. The correlation is 0.23,0.28, 0.15, 0.07, 0.26, and 0.12 for air quality stations such as Aksaray, Esenler, Kodikoy, Saryeri, Srinelver, and Sultangazi respectively. Overall, it is concluded that there is no correlation found between the PM_{10} concentrations and the total presence of smartphone users. It can be noticed that the total presence of smartphone users ranges from 0.2 million to about 0.36 million, showing daily variations in their presence across Istanbul. This variation could be due to factors like breaking in trajectory recording or smartphone users being temporarily outside the city. Future research would focus on the improvement of this FRK model by using longer-term historical air quality data and detailed individual mobility patterns with higher sampling rates. This approach would provide a clearer understanding of the relationship between air quality and human mobility. All statistical analysis was completed by using the R (version 4.3.2).

Acknowledgment

This contribution has been developed for the project “Study of mobile phone siGNals for the evalUation of the interconnections between mobility and the environment in Lombardy (SIGNUM)” CUP: F53D23010910001- PRIN 2022 PNRR M4C2 - financed by the European Union ˆa Next Generation EU (DD MUR n. 1409 del 14/09/2022)

References

Bucak, U., Arslan, T., Demirel, H., Balın, A. (2021). Analysis of strategies to reduce air pollution from vessels: A Case for the Strait of Istanbul. *Journal of Eta Maritime Science*, *9*(1), pp. 22–30. https://doi.org/10.4274/jems.2021.19327.

Finazzi, F., Paci, L. (2019). Quantifying personal exposure to air pollution from smartphone-based location data. *Biometrics*, *75*(4), 1356–1366. https://doi.org/10.1111/biom.13100.

Finazzi, F. (2016). The earthquake network project: Toward a crowdsourced smartphone-based earthquake early warning system. *Bulletin of the Seismological Society of America*, 106, pp. 1088–1099

Nguyen, H., Cressie, N., Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, *107*(499), pp. 1004–1018. https://doi.org/10.1080/01621459.2012.694717.

Robinson, T. P., Metternicht, G. (2006). Testing the performance of spatial interpolation techniques for mapping soil properties. *Computers and Electronics in Agriculture*, *50*(2), pp. 97–108. https://doi.org/10.1016/j.compag.2005.07.003.

Yurtseven, E., Vehid, S., Bosat, M., Köksal, S., Yurtseven, C. N. (2018). Assessment of ambient air pollution in Istanbul during. *Iran Journal Public Health,* 47 (8). http://ijph.tums.ac.ir.

Zammit-Mangion, A., Cressie, N. (2017). *FRK: An R Package for Spatial and Spatio-Temporal Prediction with Large Datasets*. http://arxiv.org/abs/1705.08105.

Zhu, Y., Xie, J., Huang, F., Cao, L. (2020). The mediating effect of air quality on the association between human mobility and COVID-19 infection in China. *Environmental Research*, *189*. https://doi.org/10.1016/j.envres.2020.109911.

doi.org/10.26398/asaproc.00135

The combination of expert knowledge and Bayesian networks to identify the relevant factors in diabetic kidney disease (DKD)

Debora Slanzi[ab], Irene Poli[b], Roger D. Jones[b], Gert Mayer[c]

[a] Dept of Management, Venice School of Management, Ca' Foscari University of Venice, Italy
[b] European Centre for Living Technology (ECLT), Ca' Foscari University of Venice, Italy
[c] Dept of Internal Medicine IV (Nephrology and Hypertension), Medical University Innsbruck, Austria

1. Introduction

Type 2 diabetes mellitus (T2DM) is a chronic condition characterized by elevated blood glucose levels due to insulin resistance or reduced insulin secretion. The number of adults with T2DM in Europe is expected to increase from 52.8 million in 2011 to 69 million by 2045. About 30-40% of these individuals develop diabetic kidney disease (DKD), a severe complication that decreases both the quality and duration of life and imposes a significant burden on healthcare systems. DKD is the leading cause of end-stage renal disease in developed countries (Perco et al., 2019). Initially, kidney disease in T2DM was thought to mimic that in type 1 diabetes, driven primarily by genetic predisposition and metabolic control. However, it is now understood to be more complex and multifactorial, influenced by comorbidities such as hypertension and disruptions in various biological pathways. This complexity results in significant variability in disease progression and therapy response among individuals (Galicia-Garcia et al. 2020). Understanding these mechanisms is essential for improving clinical care and developing targeted treatments (Slanzi et al., 2024, Abebe et al., 2024).

Identifying the most relevant variables that influence DKD is crucial for a more targeted approach to treatment and management (Jones et al., 2023). By focusing on these key factors, we can reduce uncertainty and improve the precision of interventions, ultimately leading to better patient outcomes. This targeted approach minimizes the need for broad-spectrum treatments, which can be less effective and carry a higher risk of side effects. Additionally, understanding the specific variables at play can help to develop personalized therapies that address each patient's unique needs, thereby enhancing treatment efficacy and slowing DKD progression.

Integrating data-derived insights with expert knowledge can significantly enhance the understanding and management of DKD. Data provides empirical evidence that uncovers patterns and relationships among variables, while expert knowledge offers contextual understanding and clinical experience. Combining these sources allows for a more comprehensive analysis, adding valuable context to raw data and improving the selection of the most impactful variables. This integration not only refines the focus of research and treatment but also ensures that the chosen variables are clinically relevant and actionable, ultimately leading to more effective and personalized healthcare strategies for patients with DKD.

The aim of this work is to develop an approach based on Bayesian networks for selecting the relevant variables that influence diabetic kidney disease (DKD). A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph. By utilizing Bayesian networks, we aim to uncover the complex relationships among various factors involved in DKD. This method allows for the integration of data and expert knowledge, providing a systematic way to identify and analyse the most impactful variables. Ultimately, this approach can lead to more precise and effective interventions for managing DKD, improving patient outcomes by focusing on the most critical aspects of the disease. The paper is structured as follows: in Sect. 2 we introduce the study conducted to achieve the data used in the analyses and the statistical approach developed to address the proposed objectives, in particular the steps of the procedure used to variable selection and how to include prior knowledge available from the literature and experts to produce more informative models;

then in Sect. 3 we present the main results achieved in analysing DKD patients highlighting the main findings of the research.

2. Materials and methods

The Data. The data utilized in this study are sourced from the PROVALID study ("PROspective cohort study in patients with type 2 diabetes mellitus for VALIDation of biomarkers"), a prospective observational study that enrolled over 4,000 T2DM patients across five European countries with normal, mild, or moderately reduced kidney function. These patients were monitored for at least four years, with annual collection of clinical data, laboratory values, and medication information. For a more detailed description of the study and the available data, refer to (Eder et al., 2018 and 2019). The outcome of the study is determined by the estimated glomerular filtration rate (eGFR) value, a marker of renal excretory capacity, recorded at each visit. The disease trajectories measured by changes in eGFR vary significantly among PROVALID participants even under stable therapy (Kerschbaum et al., 2020). This variability is likely due to differences in drug adherence, environmental factors, and heterogeneity in pathophysiology.

The number of initial variables considered in the study exceeds 120. The aim of the work is to identify which of these variables are most relevant for predicting changes in eGFR, specifically the difference in eGFR between two consecutive visits, measured as $\Delta eGFR = eGFR_{ti} - eGFR_{ti-1}$. Additionally, the study seeks to elucidate how these variables are interconnected to better understand the underlying mechanisms of the disease. By pinpointing the key variables and mapping their interrelations, the study would like to provide a comprehensive explanation of the factors driving changes in eGFR and to shed light on the complex interactions contributing to the progression of diabetic kidney disease (DKD).

After preprocessing the data to remove incomplete cases and to adjust skewed distribution through log transformation when appropriate, we include n=537 observations in our analysis. We point out that we consider datapoints without considering the longitudinal aspect because, in this initial phase, we are interested in identifying the variables responsible for a change in eGFR, independently of the timing of the visit.

Bayesian networks. To identify the network of relationships among the many factors that may be responsible for diabetic kidney disease (DKD), we use a class of probabilistic graphical models known as Bayesian networks (BNs) (Koller and Friedman, 2009; Scutari and Denise, 2014; Arora et al., 2019). BNs are directed acyclic graphs where nodes represent variables and edges denote probabilistic dependencies between these variables. Each node in the network is associated with a probability distribution that quantifies the likelihood of different values of the variable, given the values of its parent nodes.

Specifically, a BN for a set of random variables $\mathbf{X}=\{X_1,...,X_p\}$ is identified by:

– a *network structure* G, a directed acyclic graph (DAG) where nodes represent the variables **X** of the system and the directed arcs between nodes represent the probability dependences between them,

– a *set of parameters*, representing conditional probability distributions $P(X_i|Pa(X_i))$ associated to each variable X_i, i=1,...,p, where $Pa(X_i)$ are the variables that correspond to the parents of X_i in the DAG (i.e. the nodes with an arc pointing towards X_i that indicates a directed dependency with it).

The global distribution of the variables **X** is decomposed into the local distributions of the individual variables X_i as

$$P(\boldsymbol{X}) = \prod_{i=1}^{p} P(X_i|Pa(X_i))$$

The process of estimating a BN is called learning and typically involves two main steps: (1) the *structure learning* to identify the topological structure, i.e. which arcs are present in the graph and therefore which probabilistic relationships are supported by the data, and (2) the *parameter learning* to learn the conditional probability distributions that regulate the strength of the relationships.

There are many approaches in literature to estimate BNs from the data (Kitson et al., 2023): in this work we will focus on a Search & Score strategy which uses a score function in order to compare the structures of the network and then selects the structure which better fits the data. Specifically, we develop

structure learning by means of a hill-climbing search procedure and a BDe score (Koller and Friedman, 2009). BNs can model complex, multivariate relationships and are able to handle uncertainty by combining prior knowledge with observed data. They also facilitate inference and reasoning, allowing for the prediction of outcomes and the identification of causal relationships. Through these properties, BNs provide a robust framework for understanding the intricate interplay of factors contributing to DKD.

Variable selection. The Markov blanket (MB) is a crucial concept in BNs for variable selection. It represents the set of nodes that directly influence a specific node, consisting of its parents, children, and the parents of its children. This set encapsulates all the information needed to predict the target variable, effectively isolating it from the rest of the network. By focusing on the MB, one can identify the minimal and most relevant subset of variables necessary for prediction, thereby eliminating redundant or irrelevant data. This not only simplifies the model and enhances computational efficiency but also improves interpretability by highlighting the key variables that influence the outcome. Overall, using the Markov Blanket for variable selection ensures a theoretically optimal and practical approach to building efficient and comprehensible BNs (Wang et al., 2020).

Since experts in DKD are interested also in variables that indirectly impact the target of interest, it has been proposed to expand the concept of MB in a BN. This expansion includes not only the directly related variables but also those indirectly connected to the target. By doing so, the analysis can capture a broader range of influential factors, providing a more comprehensive understanding of the various elements that may affect the target outcome. This approach allows for a deeper exploration of the relationships within the network, potentially uncovering hidden or subtle influences that might otherwise be overlooked. We thus develop a variable selection procedure using the properties of BNs in several steps. Specifically, we proceed as follow:

1. BN estimation: A model averaging procedure is developed where arcs that exceed a certain threshold t are identified. This process involves performing bootstrap resampling, which means re-sampling the data k times using the bootstrap and performing structure learning separately on each of the resulting samples to collect k DAGs. The arc strength is then calculated by determining the frequency with which each arc appears in those k graphs. An "average" consensus DAG is derived by selecting those arcs that have a frequency above t. Following the literature, in this work we set the number of bootstrap replications to $k = 100$ and threshold to $t = 0.5$ (selecting only arcs with strength > 0.5) (Scutari and Nagarajan, 2013).
2. Markov blanket identification: the MB of ΔeGFR (target of the system) is selected. MB includes the target variable's parents, children, and the parents of its children, effectively encapsulating all variables that directly influence or are directly influenced by the target.
3. MB expansion: for all variables identified in the initial MB, their MB is also selected. This process iteratively expands the network to include all relevant variables and their direct and undirected relationships, ensuring a comprehensive selection of variables most pertinent to the target variable.

Additionally, expert knowledge can be integrated into the process by imposing certain known relationships in the pathophysiology of diabetic kidney disease (DKD). This ensures that domain-specific insights enhance the accuracy and relevance of the variable selection. The prior information was provided by study physicians in the form of 32 prior relationships (whitelisted arcs) derived from the pathophysiological theoretical framework as described in Slanzi et al. (2024).

3. Results

Starting with the whole set of over 120 variables, the selection process described above identify 18 variables as the most important for the study's target, ΔeGFR. The selected variables are listed in Table 1. Only two of these, CST3 and SCR, are part of the Markov blanket MB(ΔeGFR). The remaining variables constitute the MB expansion, indicating variables that appear to have undirected relationships with the target. These variables are important for eGFR and its variation because they are associated with various physiological and clinical aspects that can influence kidney function and health.

Table 1. The selected variables (in alphabetical order)

Covariate	*Description*	*Type*	*Descriptive statistics (Mean±sd / Percent%)*
ADIPOQ	Adiponectin, C1Q, And Collagen Domain Containing	Numerical	8490781± 3947052
ADMD	Age at DM2 diagnosis	Numerical	52.36 ±9.93
ADMET	Biguanides (metformin)	Categorical	True: 74% False: 26%
CST3	Cystatin C concentration in serum	Numerical	1.29±0.43
DLOOP	Loop diuretics	Categorical	True: 29% False: 71%
EGF	Epidermal growth factor concentration in urine normalized by UCREA	Numerical	8.42±0.77
EGFR	eGFR creatinine based	Numerical	64.36±18.13
FGF21	Fibroblast growth factor 21 concentration in urine normalized by UCREA	Numerical	104.62±74.39
GE	Gender	Categorical	Female: 46% Male: 54%
HB	Haemoglobin	Numerical	13.48±1.57
HDLCHOL	Serum cholesterol (HDL)	Numerical	48.04±13.60
LGALS3	Galectin 3 concentration in serum	Numerical	8824±2581.98
PHRDB	Personal history of renal disease at baseline	Categorical	Diabetic renal disease:13% Other renal disease: 12% False: 75%
SCR	Serum creatinine	Numerical	1.10±0.37
SPOT	Serum potassium	Numerical	4.54±0.49
TNFRSF1A	TNF Receptor Superfamily Member 1A concentration in serum	Numerical	2174±1020.38
UACR	Mean UACR	Numerical	2.49±1.81
UCREA	Urinary creatinine	Numerical	78.80±42.39
Target			
DEGFR	ΔeGFR= $eGFR_{ti}$- $eGFR_{ti-1}$	Numerical	-1.45±18.76

For instance, biomarkers like Adiponectin (ADIPOQ), Cystatin C (CST3), and TNF receptor superfamily member 1A (TNFRSF1A) are related to metabolic and inflammatory processes that affect renal function. Other variables like age at DM2 diagnosis (ADMD), personal history of renal disease (PHRDB), and medication usage (e.g., ADMET for metformin and DLOOP for loop diuretics) provide context on patient history and treatment, which can significantly impact kidney health. Additionally, factors such as serum creatinine (SCR), urinary creatinine (UCREA), and mean UACR provide direct measures of kidney function, while variables like haemoglobin (HB) and serum potassium (SPOT) can indicate related systemic health conditions. These variables collectively help in understanding and predicting changes in eGFR, providing a comprehensive overview of factors influencing kidney function.

To understand the relevance of the selected variables for the target, we estimate a Bayesian network using p_1=18+1=19 variables (the selected variables plus the target, ΔeGFR) to evaluate the strength of their relationships. Additionally, since it is possible to incorporate expert prior information, we also estimate a Bayesian network by introducing prior knowledge from the literature and practical experience in the field. This is accomplished using whitelisted arcs, which represent well-known dependencies that should be enforced in the graph, based on 32 prior relationships derived from the pathophysiological theoretical framework constructed on 17 specific variables (Slanzi et al. 2024). As many of these prior relationships involved variables not included in the selected set, it was necessary to expand the list of variables to encompass all relevant ones, resulting in a new set of p_2=30 variables (only 7 of these are in the set of 18 selected). Subsequently, we estimate the BNs as described in step 1 of the proposed procedure outlined in Section 2. In Table 2, we present several measures of graphical differences: the number of nodes (Num. nodes), the number of arcs (Num. arcs), the average Markov blanket size (Av. MB size), the average neighbourhood size (Av. neighb. size), and the Bayesian information criterion (BIC) to evaluate the model fit to the data.

Table 2. Measures of graphical differences. *Original* refers to the BN estimated using the initial complete set of variables, *Selected* refers to the BN learned using the set of selected variables only, *Selected + Prior* refers to the BN learned by using the set of selected variables and imposing expert knowledge.

	Num. nodes	Num.arcs	Av. MB size	Av. neighb. size	BIC	Num. MB(ΔeGFR)
Original	123	199	5.14	3.24	-137443.4	2
Selected	19	43	6.74	4.53	-35847.18	6
Selected+Prior	30	105	13.00	7.00	-55796.59	6

Additionally, we highlight the cardinality of the MB of the target variable (Num. MB(ΔeGFR)). From the results, we can see that the "selected" BN has fewer arcs, indicating that the data suggest relationships considered to be more robust. Adding prior relationships enriches the set of arcs, making the network denser. Notably, the number of variables forming the MB of the target is larger compared to the original network, highlighting that reducing the number of variables to those that are truly relevant can lead to more accurate results.

To evaluate the effectiveness of the proposed procedure in identifying an efficient set of variables that are useful for reducing the noise from redundant variables and deriving the pattern of DKD pathophysiology with clinically relevant and actionable variables, we calculate the predictive accuracy of the BNs in terms of the correlation between the observed and predicted values for the target variable. This predictive accuracy is assessed using 10-fold cross-validation. We compute the correlation between the observed and predicted values for ΔeGFR and this quantity is called predictive correlation. The procedure is repeated M=30 times to achieve the mean posterior predictive correlation for the three BNs as previously described. The results are presented in Figure 1. We find that changes in eGFR are more accurately predicted when using the reduced set of variables. Furthermore, using prior information improves the prediction accuracy, although not substantially, demonstrating how the expertise of clinicians can enhance the description of the system.

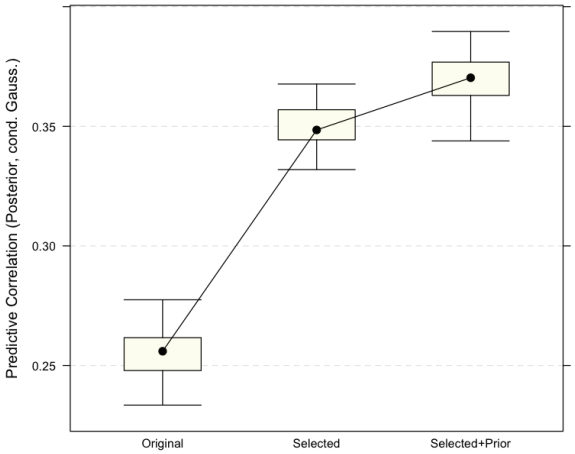


Figure 1. Predictive correlation of ΔeGFR. *Original* refers to the BN estimated using the initial complete set of variables, *Selected* refers to the BN learned using the set of selected variables only, *Selected + Prior* refers to the BN learned by using the set of selected variables and imposing expert knowledge.

By incorporating expert knowledge in the modelling phase, we demonstrate that the synergy between data and expert prior information is a great source of valuable assistance in gaining new insights into studying complex disorders and understanding DKD with respect to a predefined target of interest. In fact, our results indicate that more than half of the variables identified as relevant to the target of interest are not part of the a priori pathophysiology theoretical framework identified by clinical experts. This has provided new evidence regarding the pathophysiology of the disease, highlighting how statistical methods and data-driven technologies can support expert decisions with new knowledge. Future research will consider the longitudinal dimension of the data to assess the model's ability to capture temporal dynamics,

and comparisons with other methods for variable selection.

Acknowledgments

This work is supported by the project DC-ren, that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 848011.

References

Abebe, S., Poli, I., Jones, R.D., Slanzi, D. (2024) Learning optimal dynamic treatment regime from observational clinical data through reinforcement learning. *Machine Learning & Knowledge. Extraction*, **6**, pp. 1798–1817.

Arora, P., Boyne, D., Slater, J.J., Gupta, A., Brenner, D.R., Druzdzel, M.J. (2019) Bayesian networks for risk prediction using real-world data: A tool for precision medicine. *Value in Health,* **22**(4), pp. 439–445.

Eder, S., Leierer, J., Kerschbaum, J., Rosivall, L., Wiecek, A., de Zeeuw, D., Mark, P.B., Heinze, G., Rossing, P., Heerspink, H.L., Mayer, G. (2018) A prospective cohort study in patients with type 2 diabetes mellitus for validation of biomarkers (PROVALID) - Study design and baseline characteristics, *Kidney Blood Press Research*, **43**(1), pp. 181–190.

Galicia-Garcia, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K.B., Ostolaza, H., Martin, C. (2020) Pathophysiology of type 2 diabetes mellitus, *International Journal of Molecular Sciences*, **21**(17), 6275.

Eder, S., Leierer, J., Kerschbaum, J., Rosivall, L., Wiecek, A., de Zeeuw, D., Mark, P.B., Heinze, G., Rossing, P., Heerspink, H.L., Mayer, G. (2019) Guidelines and clinical practice at the primary level of healthcare in patients with type 2 diabetes mellitus with and without kidney disease in five European countries, *Diabetes & Vascular Disease Research*, **16**(1), pp. 47–56.

Jones, R.D, Abebe, S., Distefano, V., Mayer, G., Poli, I., Silvestri, C., Slanzi, D. (2023) Candidate composite biomarker to inform drug treatments for diabetic kidney disease. *Frontieres in Medicine*, **10**, 1271407

Kerschbaum, J., Rudnicki, M., Dzien, A., Dzien-Bischinger, C., Winner, H., Heerspink, H.L., Rosivall, L., Wiecek, A., Mark, P.B., Eder, S., Denicolò, S., Mayer, G. (2020) Intra-individual variability of eGFR trajectories in early diabetic kidney disease and lack of performance of prognostic biomarkers, *Scientific Reports*, **10**, 19743.

Kitson, N.K., Constantinou, A.C., Guo, Z., Liu, Y., Chobtham, K.(2023) A survey of Bayesian network structure learning, *Artificial Intelligent Review*, **56**, pp. 8721-8814.

Koller, D., Friedman, N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Perco, P., Pena, M., Heerspink, H.J.L., Mayer, G. (2019) Multimarker panels in diabetic kidney disease: The way to improved clinical trial design and clinical practice?, *Kidney International Reports*, **4**(2), pp. 212–221.

Scutari, M., Denis, J.B. (2021) *Bayesian Networks with Examples in R*. Chapman & Hall.

Scutari, M., Auconi, P., Caldarelli, G., Franchi, L. (2017) Bayesian networks analysis of malocclusion data, *Scientific Report*, **7**(1), 15236.

Scutari, M., Nagarajan, R. (2013) On Identifying Significant Edges in Graphical Models of Molecular Networks. *Artificial Intelligence in Medicine*, **57**(3), pp. 207–217.

Slanzi, D., Silvestri, C., Poli, I., Mayer, G. (2024). Exploiting the potential of Bayesian networks in deriving new insight into diabetic kidney disease (DKD), in *Artificial Life and Evolutionary Computation. WIVACE 2023*, eds M. Villani, S. Cagnoni, R. Serra, R. *Communications in Computer and Information Science*, vol 1977. Springer, Cham.

Wang, H., Ling, Z., Yu, k., Wu, X. (2020) Towards efficient and effective discovery of Markov blankets for feature selection, *Information Sciences*, **509**, pp. 227-242.

doi.org/10.26398/asaproc.00136

Understanding consumer behavior in the plant-based meat market: Insights from Bayesian network analysis

Debora Slanzi[a], Pietro Lanzini[a], Francesca Checchinato[a], Cinzia Colapinto[a]
[a] Venice School of Management, Ca' Foscari University of Venice, Italy

1. Introduction

In recent years, the food industry witnessed a growing interest in meat analogues and plant-based meat (PBM), which aim to mimic traditional meat in terms of taste, texture, and appearance. While it's clear that these new products target vegetarian and vegan consumers, PBM also offers an alternative for those concerned with health and ethical considerations, due to its potentially lower environmental impact in terms of emissions, water and land use compared to conventional meat.

Despite the increasing market presence, PBM remains a niche product, with several aspects limiting its broader adoption. These include higher prices, limited availability in stores, skepticism about its resemblance to real meat, and a phenomenon known as food neophobia (Barrena and Sánchez, 2013). Understanding the factors influencing PBM purchase decisions is therefore essential for companies willing to develop effective strategies.

In this work, we contribute to the literature on PBM consumer behavior conducting an online survey on a sample of mainly Italian consumers (n=988). We analyzed the data using Bayesian networks (BNs), a class of probabilistic graphical models identifying significant relationships among variables representing the system under study. A BN is able to represent a set of variables and their conditional dependencies via a directed acyclic graph: by utilizing BN, we, hence, aim to uncover potential relationships among factors representing the relevant constructs of the literature explaining PBM purchase.

The paper is structured as follows: in Sect. 2 we introduce our theoretical framework, and the methodological approach used to derive the network of relations among collected data; then in Sect. 3 we illustrate our main findings, and we discuss avenues for future research.

2. Materials and methods

Theoretical framework. We introduce the theoretical framework underlying the literature on consumer behavior regarding PBM, describing the relevant constructs used in the survey. Pertinent constructs have been selected relying on mainstream theoretical frameworks, assuming that behaviors are either the outcome of a cognitive rational process, such as in the case of the theory of planned behavior (TPB, Ajzen 1985) and the norm activation model (NAM, Schwartz 1977), or of automatic responses to recurrent situations, such as in habit-based theories (Verplanken 2018) or, typically, on a combination of both.

According to TPB, intentions represent the closest antecedent of behaviors and are in turn determined by three predictors: attitudes (i.e., the generic predisposition, either positive or negative, towards the activity), subjective norms (i.e., expectations of relevant ones about adopting a particular behavior, hence reflecting social pressure) and perceived behavioral control (i.e., measure of how individuals perceive a specific task to be easy or difficult to perform, since not all behaviors are under volitional control). TPB represents one of the most popular models for investigating consumer behavior and has been extensively adopted in a wide range of behavioral domains, including PBM purchase (Chen 2024, Wang and Scrimgeour 2021). NAM on the other hand focuses on altruistic drivers and assumes that the closest predictor of behavioral intentions and actual behaviors is represented by personal norms, which are feelings of moral obligations

towards performing an activity (or refraining from doing so): in turn, personal norms depend on awareness of consequences and ascription of responsibility. Although to a lesser extent compared to TPB, also NAM has been adopted to investigate PBM consumption (Xueyun et al. 2024). Thirdly, habit theories suggest focusing on the role exerted by habits: frequently repeated behaviors that have become automatic responses to stable and regularly encountered contexts might shape dietary patterns over time (Kurz et al. 2015). Food purchase and consumption can by nature be regarded as activities where developing a habit is common, hence there is growing literature adopting such construct to analyze and explain PBM purchases (Seffen and Dohle 2023): habit theories emphasize that once a food consumption pattern becomes habitual, it is resistant to changes, such as a shift to PBM. The purpose of this paper is multi-dimensional. First, we aim to verify whether the hypotheses of the theoretical framework are supported and to provide insights into new potential relationships emerging from the data, which could generate further hypotheses. Next, we evaluate the strength of the relationships identified among the factors representing key constructs in the extant literature on PBM and explain their impact on the intention to adopt PBM.

Data. A questionnaire was developed and administered online to investigate consumer attitudes and behaviors using constructs from the previously mentioned theories. After a preprocessing phase, the sample comprised 988 complete responses. The sample consists of 65% women and 35% men, with an average age of 24 years, and most respondents (89%) are of Italian nationality. Additionally, respondents reported following various diets: 84% omnivorous, 13% pescatarian or flexitarian, and only 3% vegetarian or vegan. Among the sample, 42% already tried PBM. A subset of items was selected to measure the relevant constructs of the hypothesized theoretical framework. Specifically, three items were identified for resistance to change (RTC_i), four items for Habits ($Habit_1$=frequency; $Habit_2$=automaticity; $Habit3$=doing without thinking; $Habit_4$=idiosyncrasy), and four items from the theory of planned behavior (*Att*=Attitudes, *SN*=Subjective Norms, PBC_i=Perceived Behavioral Control). Additionally, one item related to the norm activation model (*NAM*) was included, representing personals (*PN*). The main target of interest for the study was measured with an item assessing the intention to include PBM in one's diet (*Intention*). All these variables were measured using a 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree. Figure 1 presents some statistics of the observed variables.

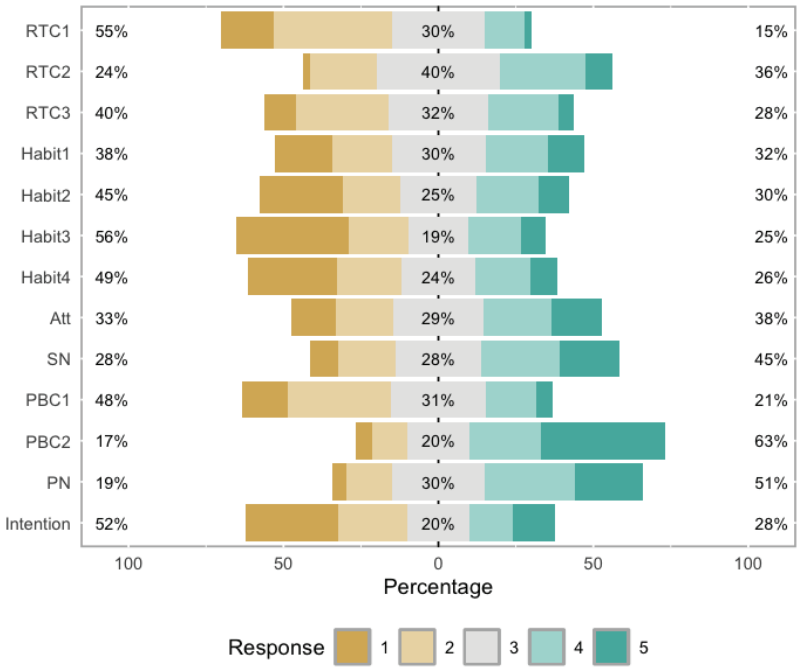


Figure1. Descriptive statistics of the items used in the analysis

Bayesian networks.

BNs are directed acyclic graphs where nodes represent variables and arcs denote probabilistic dependencies between these variables (Koller and Friedman, 2009; Scutari and Denise, 2021; Briganti et al., 2024). Each node in the network is associated with a probability distribution that quantifies the likelihood of different values of the variable, given the values of its parent nodes.

Specifically, a BN for a set of random variables $\mathbf{X}=\{X_1,...,X_p\}$ is identified by:

- a *network structure* G, a directed acyclic graph (DAG) where nodes represent the variables **X** of the system and the directed arcs between nodes represent the probability dependences between them,
- a *set of parameters*, representing conditional probability distributions $P(X_i|Pa(X_i))$ associated to each variable X_i, i=1,...,p, where $Pa(X_i)$ are the variables that correspond to the parents of X_i in the DAG (i.e. the nodes with an arc pointing towards X_i).

The global joint distribution of the variables **X** is decomposed into the local conditional distributions of the individual variables X_i

$$P(\boldsymbol{X}) = \prod_{i=1}^{p} P(X_i|Pa(X_i))$$

The process of estimating a BN is called learning and typically involves two main steps: (1) the structure learning to identify the network structure, i.e. which arcs are present in the graph and therefore which probabilistic relationships are supported by the data, and (2) the parameter learning to learn the conditional probability distributions that regulate the strength of the relationships. When all the variables are continuous or is treated as approximately continuous as common practice with Likert scales, we refer to Gaussian Bayesian networks (GBNs). In a GBN every node follows a normal distribution, and the joint probability distribution is expressed as a multivariate normal (Scutari and Denise, 2014, Briganti et al., 2024).

There are many approaches in literature to estimate BNs from the data (Kitson et al., 2023): in this work we will focus on a Search & Score strategy which uses a score function in order to compare the structures of the network and then selects the structure which better fits the data. Specifically, we develop structure learning by means of hill-climbing search procedure and a BDe score (Koller and Friedman, 2009). Furthermore, to reduce the impact of the noise present in the data, model averaging learning techniques can be used to improve the reliability of structure learning (Scutari and Nagarajan, 2013). The process consists of performing bootstrap resampling, i.e. re-sample the data *k* times using bootstrap and perform structure learning separately on each of the resulting samples, thus collecting *k* DAGs. Then the arc strength is calculated, i.e. the frequency with which each arc appears in those *k* graphs deriving an "average" consensus DAG by selecting those arcs that have a frequency above a certain threshold *t*. We bootstrapped the learning of the BN *k*=1,000 times and included in the final network the edges that appeared in more than 80% of the bootstrapped networks, determining their direction based on those present in more than 50% of the bootstrapped networks. In this way, all arcs are robust and the strength of all the estimated relationships can be quantified. We propose using Bayesian Networks (BNs) over traditional consumer behaviour approaches (such as PLS or regression analysis), as they allow for a deeper exploration of complex inter-variable relationships beyond simply testing theoretical models from the literature (Briganti et al., 2024), while also relaxing some of the assumptions typical of latent factor models. This could provide insights on new potential relationships emerged from the data. Moreover, this approach enables not only the estimation of dependencies among items but also assesses the robustness of relationships in terms of directionality and, consequently, potential effects in explaining the PBM market.

3. Results

The BN learned from the survey data is shown in Figure 2. All the relationships have an arc strength above 0.80, meaning they were estimated in over 80% of the 1,000 bootstrapped DAGs. Most of them have a strength of one (e.g., Att→SN, Att→PN, Att→Intention, all interconnections among Habit items, and several others), indicating the robustness of the learned relationships. The network structure reveals that the variables representing measurements of the same latent construct are directly connected, indicating that the approach effectively identifies items associated with the same theoretical framework. Furthermore, the analysis confirms TPB as Attitudes, Subjective norms (SN), and Perceived behavioral control (PBC) all have an impact on Intention. Specifically, while Attitudes (Att = "How pleasant would it be for you to include PBM in your diet?") and PBC (PBC1="How easy is it to find PBM products at the supermarkets you know?") have a direct relationship with Intention ("How likely will you include PBM in your diet over the next month?"), Subjective norms (SN= "Most people who are important to you approve if you include PBM in your diet") have an indirect relationship mediated by PBC. This is in line with the findings of Contini et al. (2020), which emphasize the role of market availability – described as PBC_1 - as a predictor of intention. On the other hand, in previous research SN usually directly impact intention. In our model SN is mediated by market availability. The reason could be underlined by the novelty of the products. It appears that consumers are externally driven when it comes to a novel product.

In other words, a consumer will consider buying PBM only if they perceive that significant others approve of their choice to consume it. At that point, he/she considers the availability of the product, that is still limited in the store. This suggests that SN are not enough to drive the intention to buy PBM. As far as PN are concerned, they are influenced by meat consumption habits.

Habits ("Purchasing meat is something" Habit1="I do frequently"; Habit2="I do automatically"; Habit3="I do without thinking"; Habit4="that is typically me") do not have a direct impact on Intention; however, they exert a direct influence from Habit4 on Attitudes and Personal norms (PN="When choosing which food products to purchase, I feel that it is my duty to take into consideration the environmental and social consequences of this choice as well.").

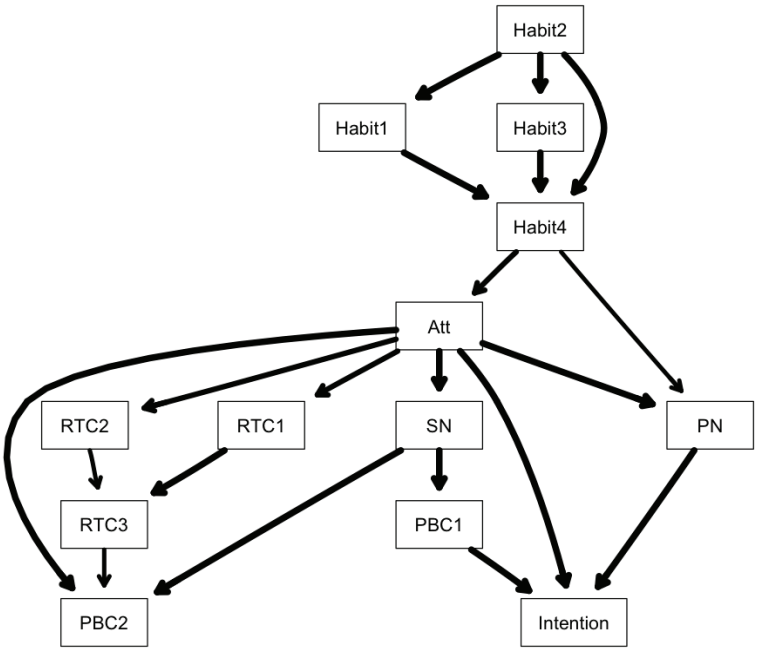


Figure 2. The Bayesian Network

Resistance to change (RTC1="I like to do the same old things rather than try new and different ones"; RTC2="I often change my mind"; RTC3="Often I feel a bit uncomfortable even about changes that may potentially improve my life") appears to be marginal in relation to Intention, but RTC1 is directly influenced by Attitudes and surprisingly does not affect Habits.

Focusing on the conditional Gaussian distribution of Intention, we derive the following regression equation:

$$\text{Intention} = -0.644 + 0.747 \cdot \text{Att} + 0.119 \cdot \text{PBC1} + 0.180 \cdot \text{PN}$$

From this equation, we observe that for each one-unit increase in Att, Intention rises by approximately 0.747, holding other variables constant, suggesting a strong positive influence of attitudes on PBM purchase intention.

Additionally, both PBC1 and PN have a positive effect on Intention, but smaller than that of Attitudes.

We investigated the construct from relevant theoretical frameworks, specifically from TPB, NAM and habit theories. Our goal was to provide a blueprint of questions answerable by using directed graphs in survey data. The BN for intention to include PBM supports hypotheses emphasizing key factors, such as Attitudes and Perceived behavioral control, as relevant in shaping this intention. Results suggest that, while resistance to change plays a marginal role in orienting behaviors, past habits affect attitudes towards PBM which in turn (along with perceived behavioral control) are direct antecedents of purchasing intention. Social pressure, on the other hand, has only an indirect effect on the latter. The analysis helped researchers identify the plausible new relationships in the survey data, gaining insight on the relevant constructs and providing evidence in generating new hypotheses. Future research will focus on investigating potentially causal relationships among the relevant factors identified for the intention to consume PBM and purchase behaviour, including an analysis of differences between consumers with and without prior experience

References

Ajzen, I. (1985) From intentions to actions: A theory of planned behavior, in *Action-Control: From Cognition to Behavior,* eds J. Kuhl, J. Beckmann, Springer, Heidelberg, pp. 11-39.

Barrena, R., Sánchez, M. (2013) Neophobia, personal consumer values and novel food acceptance. *Food Quality and Preference,* **27**(1), pp. 72-84.

Briganti,G., Decety, J., Scutari, M., McNally, R.J., Linkowski, P. (2024) Using Bayesian networks to investigate psychological constructs: The case of empathy. *Psychological Reports*, 2022 (OnlineFirst).

Chen, M.F. (2024) Consumer food choice motives and willingness to try plant-based meat: moderating effect of meat attachment. *British Food Journal*, **126**(3), pp. 1301-1324.

Contini, C., Boncinelli, F., Marone, E., Scozzafava, G., Casini, L. (2020) Drivers of plant-based convenience foods consumption: Results of a multicomponent extension of the theory of planned behaviour. *Food Quality and Preference*, **84**, 103931.

Kitson, N.K., Constantinou, A.C., Guo, Z., Liu, Y., Chobtham, K. (2023) A survey of Bayesian network structure learning, *Artificial Intelligent Review*, **56**, pp. 8721-8814.

Koller, D., Friedman, N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Kurz, T., Gardner, B., Verplanken, B., Abraham, C. (2015) Habitual behaviors or patterns of practice? Explaining and changing repetitive climate-relevant actions. *Wiley Interdisciplinary Reviews: Climate Change*, **6**(1), pp. 113-128.

Schwartz, S.H. (1977) Normative influences on altruism. *Advances in Experimental Social*

Psychology, **10**, pp. 221-279.
Scutari, M., Denis, J.B. (2021) *Bayesian Networks with Examples in R (2nd ed).* Chapman & Hall.
Scutari, M., Nagarajan, R. (2013) On identifying significant edges in graphical models of molecular networks. *Artificial Intelligence in Medicine*, **57**, pp. 207–217.
Seffen, A.E., Dohle, S. (2023) What motivates German consumers to reduce their meat consumption? Identifying relevant beliefs. *Appetite*, **187**, 106593.
Verplanken, B (2018). *The Psychology of Habit: Theory, Mechanisms, Change, and Contexts*. Springer
Wang, O., Scrimgeour, F. (2021) Willingness to adopt a more plant-based diet in China and New Zealand: Applying the theories of planned behaviour, meat attachment and food choice motives. *Food Quality and Preference*, **93**, 104294.
Xueyun, Z., Mamun, A.A., Hayat, N. (2024) Modeling the significance of knowledge, beliefs, and norms on intention to consume plant-based meat alternatives in China. *Food Ethics*, **9**(2), pp. 1-21.

doi.org/10.26398/asaproc.00137

Exploring childfree and childless on social media: A text analysis approach

Agostino Stavolo [a,b], Giulia Cavrini [a], Elisa Cisotto [a], Marina Marino [b]
[a] Faculty of Education, Free University of Bozen, Bozen, Italy
[b] Department of Social Sciences, University of Naples Federico II, Naples, Italy

1. Introduction

Since the early 1960s, Italy's fertility rate has undergone a significant decline, particularly from the 1970s onward. By the 1990s, the rate had dropped well below the replacement level—a threshold needed to maintain population stability—and it has remained among the lowest in Europe and OECD countries (Zambon et al., 2020). The prolonged low fertility, combined with increasing longevity, has led to an aging population, with fewer young and working-age individuals and a growing number of older adults. This demographic imbalance poses significant socio-economic challenges; for example, fewer active workers are available to support a rising number of retirees, increasing pressure on pension and healthcare systems. The decline in fertility rates is shaped by both voluntary and involuntary factors and also reflects changes in social norms around parenthood. Notably, there has been a rise in people identifying as either childless or childfree, representing two distinct pathways to non-parenthood (Neal, 2021; Blackstone, 2014). The term *childless* suggests a lack of something desired—referring to people who want children but do not have them, often due to infertility, lack of a partner, or financial constraints. By contrast, *childfree* refers to individuals who actively choose not to have children, either from the outset or after a period of being childless. This term carries a positive connotation, emphasizing a sense of freedom from the responsibilities associated with parenthood. This preliminary study aims to explore how these evolving identities—*childless* and *childfree*—are represented in the Italian press. Using the topic modeling technique latent Dirichlet allocation (LDA), we analyzed Italian online newspaper articles to identify key themes and trace their evolution over time, providing insights into how these narratives have shifted.

2. Literature review

Since the 1960s, Italy has experienced profound socioeconomic and cultural transformations that have substantially influenced fertility patterns. This period saw a significant decline in birth rates, with a particularly sharp reduction at the national level beginning in the mid-1970s. While a modest recovery was observed around 2010, fertility rates have remained persistently below replacement levels, reflecting broader demographic trends across Western nations. This decline reflects not only demographic shifts but also evolving social norms, with both *childfree* and *childless* identities becoming increasingly visible. While *childless* implies a desire for children that remains unfulfilled due to external factors, *childfree* describes individuals who choose not to have children. However, the choice to remain childfree is not always fixed and may change over time as individuals reassess their priorities in different social contexts (Moore, 2017). For many, being childfree stems from personal fulfillment or concerns about society and the environment, but this choice often faces social stigma. Childfree individuals are sometimes labeled as selfish or accused of depriving society of future contributors (Palmer, 2019; Blackstone and Stewart, 2012). In recent decades, however, acceptance of diverse family structures has grown, encompassing adoption, surrogacy, single parenting, cohabitation, grandparent guardianship, and LGBTQIA+ partnerships (Bergstrom-Lynch, 2020). Interest in childfree identities has also expanded in both media and academia. Platforms like *Time* magazine and the Reddit community r/childfree, with over 625,000 members, provide spaces where childfree individuals can share experiences, counter stigma, and

assert values like autonomy and intentionality, challenging societal assumptions about parenthood (Hintz and Brown, 2019; Morison et al., 2016). Additionally, the study by Hintz and Haywood (2021) analyzes media framing of voluntary childlessness in 103 American news articles from 1989 to 2018, identifying five dominant frames, seven-character categories, and three domains connecting childlessness to demographic, political, and cultural changes.

3. Methodological strategy

The paper aims to identify the main themes expressed by articles on Google and to track their trends over time to understand the primary narratives in the Italian press and their evolution. Developing a Python code, we extracted a total of 1,067 articles from Google with the queries "childfree", "childless", and "senza figli" for the years 2013 (127 articles), 2019 (156 articles), 2020 (186 articles), 2021 (189 articles), 2022 (225 articles), and 2023 (184 articles). Specifically, we started our preliminary analysis from 2013 to ensure we had journal articles spanning a 10-year period up to our most recent year of study. So, we analyse trends across three key periods: before, during, and after the COVID-19 pandemic. After extraction, a series of cleaning and pre-processing steps were applied to convert unstructured data into structured form. We parsed and tokenized each text, breaking it down into distinct tokens (words or strings) separated by spaces, punctuation, or special characters. This decomposition followed a bag-of-words approach, treating the text as a multi-set of words, without accounting for word order or grammar role. Subsequently, we reduce the linguistic variability according to normalization techniques such as correcting misspellings, standardizing accents, removing numbers, and converting text to lowercase. Grammar tagging identified nouns, adjectives, and verbs, which were then lemmatized to their canonical forms (e.g., verbs in the present infinitive, nouns in the masculine singular). The tokens were compiled into a vocabulary list containing distinct words (types) and their total occurrences across the text. To enhance relevance, we remove stop-words (e.g., prepositions, conjunctions) and terms with a frequency of fewer than three occurrences (Misuraca and Spano, 2020). We had a final vocabulary of 2,980 types and 15,458 tokens across 1,067 documents. Following the preprocessing steps, we applied a topic modeling techniques, known as LDA. LDA is a generative probabilistic model used in unsupervised machine learning algorithms (Blei et al. 2003) where the documents are represented as random mixtures of latent topics, where each topic is characterized by a distribution of words. The Dirichlet distribution determines the distribution of topics for each document. In the generative process, the results of the Dirichlet distribution are used to assign words in a document to various topics.

4. Preliminary results

In Figure 1, we assess the quality of three topic modeling techniques - latent Dirichlet allocation (LDA), latent semantic analysis (LSA), and non-negative matrix factorization (NMF) - to support our chosen model. To compare the quality of topics extracted by each technique, we apply a coherence measure (Röder et al., 2015). Topic coherence is quantified by calculating the average or median of the pairwise word similarities among the top words of a given topic, controlling for statistical independence. This involves computing the co-occurrence probability of word pairs and comparing these probabilities to their individual occurrences using Pointwise Mutual Information (PMI). A high coherence score indicates that words within a topic frequently appear together in documents, suggesting that the topic has semantic relevance. We use all terms in the vocabulary to define the coherence measure. Based on this evaluation, LDA outperformed the other techniques across most topics, especially for the range 5-25.

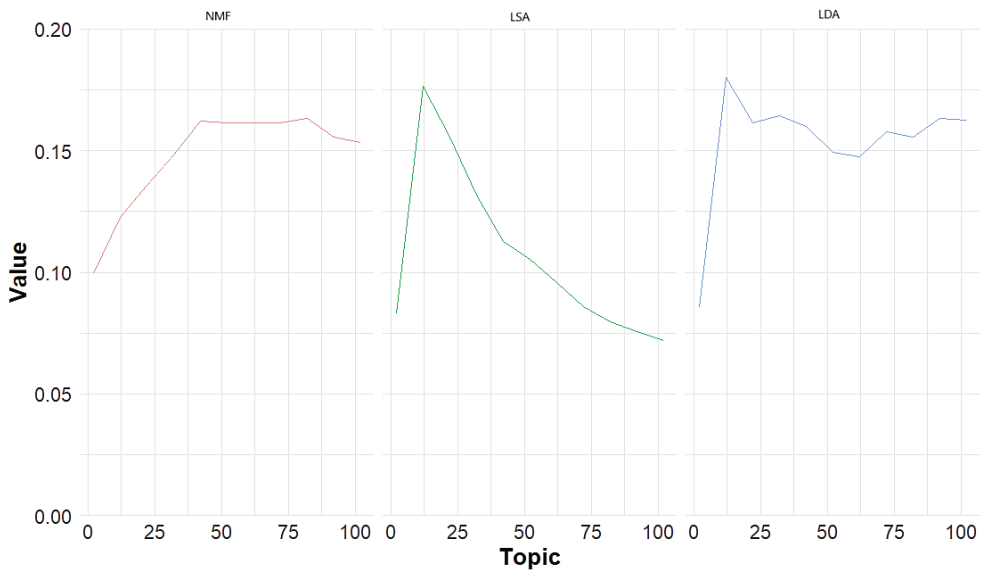


Figure 1 – Comparison of NMF, LSA and LDA according to coherence measure

Specifically, to determine the optimal number of topics, we chose the shortest distance available between metrics, as a reverse ratio of dissimilarity (De Mascellis et al., 2022). The number of topics we determined that both minimize Arun (2010) and Cao Juan (2009) and maximize Griffiths (2004) and Deveaud (2014) is $k = 5$, as shown in Figure 2.

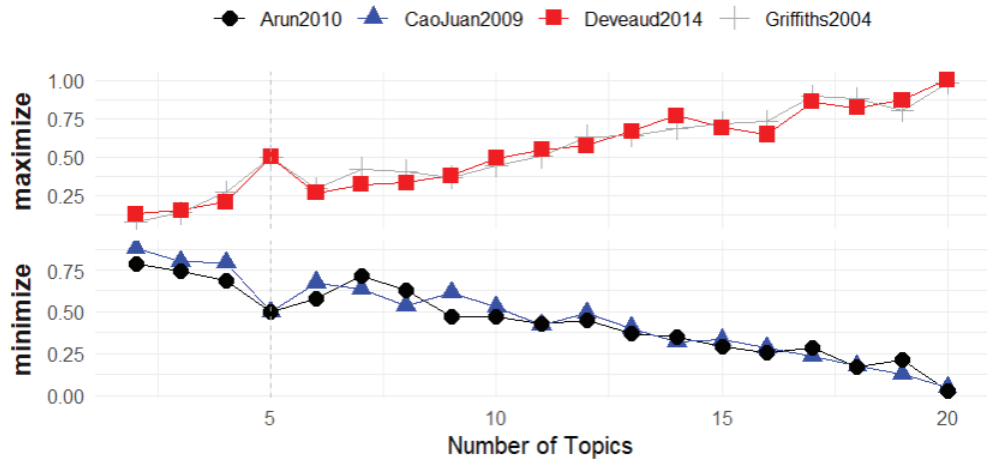


Figure 2 – Selection of *k* topics

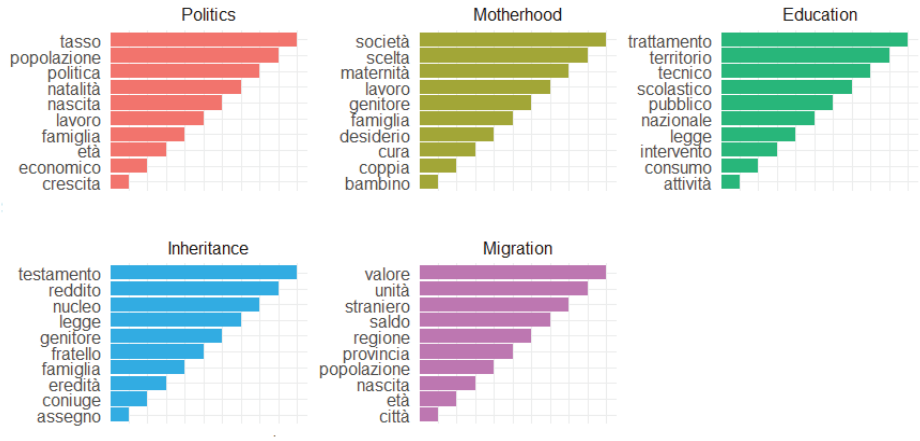


Figure 3 – Topics by latent Dirichlet allocation

In Figure 3, we reported the top ten terms associated with the extracted topics by LDA. Political themes focus on demographic dynamics and societal changes, emphasizing the need for policies that address population trends and economic challenges. These discussions reflect a broader debate between individual freedom and collective societal expectations, particularly regarding personal choices and their implications for societal stability. The political sphere often seeks to balance these tensions through interventions aimed at fostering inclusivity and mitigating disparities. The theme of motherhood underscores the complex interplay between societal expectations and personal choices, highlighting how cultural norms influence perceptions of family and parental roles. It brings to light the dual pressures of conforming to traditional ideals while also navigating individual aspirations. The discourse around motherhood reflects the broader challenges of balancing personal autonomy with collective cultural values. Topic of Education emerges as a cornerstone of societal development, representing both an investment in future generations and an arena for personal and professional growth. It reflects tensions between inclusivity, accessibility, and individual goals, revealing how education shapes societal structures while also serving as a pathway for personal fulfillment and potential. Inheritance highlights economic and legal dimensions of resource distribution, reflecting broader societal values around legacy and family structures. This theme delves into how individuals navigate complex familial and societal expectations, often balancing personal priorities with cultural norms surrounding wealth and responsibility.

Italian newspapers have extensively covered the migration topic, addressing issues related to demographic shifts, population movements, and the impacts on cities and regions. This topic includes the integration of immigrants into society, the economic and social consequences of migration, and the policies at both national and local levels. The importance of migration is often highlighted in the context of combating the crisis of declining birth rates and fertility rates. Migration is seen as a potential solution to rejuvenate the workforce and maintain population levels, thereby supporting economic stability and growth. The topic related to the demography examines demographic trends and their far-reaching implications. Discussions focus on birth rates, population growth, and the economic and social policies affecting these trends. This topic considers the broader impacts of demographic changes on various sectors of society, including labor markets, healthcare systems, and economic growth. The changing age structure of the population, the challenges of an aging society, and the potential benefits of demographic diversity are also explored.

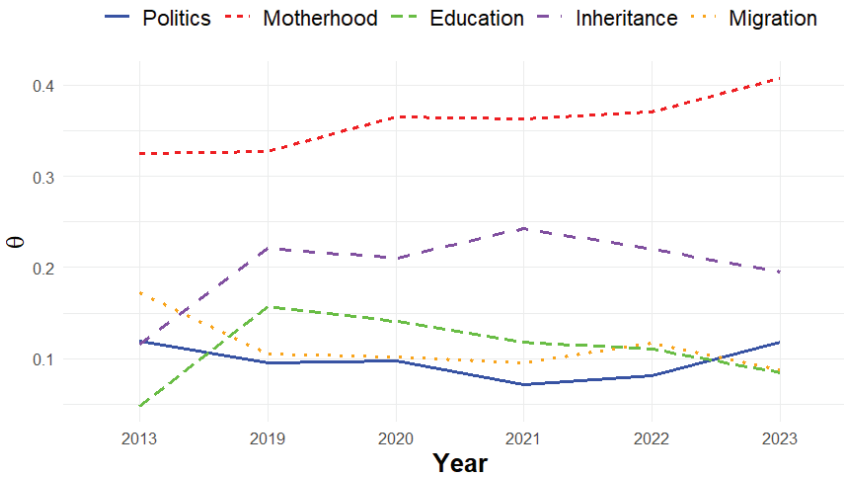


Figure 4 - Topic proportions over time

Figure 4 illustrates the evolution of discussions related to being childfree and childless from 2013 to 2023, with each line representing a different topic and its relative prominence over time. Politics (blue line) related to motherhood emerges as the most dominant and steadily growing topic, highlighting its increasing importance in public discourse. This growth likely reflects the foundational role of motherhood within social structures, explaining its consistent presence in discussions. Conversations about motherhood (red line) often revolve around societal expectations placed on women, the evolving role of family in contemporary society, and shifting definitions of parenthood in a world where childlessness is increasingly accepted as a valid lifestyle choice. Similarly, education (green line) remains a cornerstone of societal discourse, encompassing debates around access, quality, and its role in shaping future generations. The stability of this topic over time underscores its ongoing relevance, with periodic fluctuations that may be tied to specific contextual events or policy changes. In contrast, discussions around inheritance (purple line) and migration (orange line) show a noticeable decline over the observed period. Inheritance, particularly in the last two years, appears to have diminished in prominence within media narratives, perhaps due to shifting priorities or reduced focus on wealth transfer and legacy management. Migration also exhibits a downward trend, likely influenced by changes in policies and governance that have made this topic less central to discussions. Notably, migration, which was once debated in terms of its potential to counteract declining fertility rates, now receives less attention in this context.

5. Conclusions

The analysis of Italian newspapers provides significant insights into the evolving societal discourse and the main themes covered by the press. The increasing discussions around parenthood and demography indicate a heightened awareness and concern about these topics, reflecting their critical importance in societal roles, personal choices, and work-life balance responsibilities. The steady coverage of policy, migration, and inheritance topics underscores their ongoing relevance and influence on societal structures and individual lives. A detailed examination of trends offers a comprehensive understanding of how priorities and challenges within Italian society have shifted over time. By utilizing latent Dirichlet - allocation (LDA) for topic modeling, the study effectively highlights the primary issues and narratives in the media, providing a robust foundation for formulating targeted, evidence-based recommendations to support policies and initiatives. The analysis specifically focuses on articles that explicitly use the terms childless and childfree, enabling a more nuanced exploration of how these identities are represented and contrasted in the media. The thematic analysis of related articles reveals that childless narratives often emphasize societal and structural barriers, while childfree narratives highlight individual agency and the rejection of traditional norms. This distinction underscores the importance of differentiating between

involuntary and voluntary non-parenthood in both media representations and policy discussions.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining:* 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India. Proceedings. Part I, pp. 391-402.

Bergstrom-Lynch, C. (2020). Free to be you and me, maybe: Lesbian, gay, bisexual, and transgender parents doing gender with their children. *Journal of Gender Studies*, **29**(3), pp. 282-294.

Blackstone, A. (2014). Childless… or childfree? *Contexts*, **13**(4), pp. 68-70.

Blackstone, A., Stewart, M. D. (2012). Choosing to be childfree: Research on the decision not to parent. *Sociology Compass*, **6**(9), pp. 718-727.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, pp. 993-1022.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S. (2009). A density-based method for adaptive LDA model selection. N*eurocomputing*, **72**(7-9), pp. 1775-1781.

De Mascellis, A. M., Grassia, M. G., Stavolo, A. (2022). A. topic modeling and sentiment analysis: An analysis of tweets about the game cyberpunk 2077. *In JADT 2022: 16th International Conference on Statistical Analysis of Textual Data*.

Deveaud, R., SanJuan, E., Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, **17**(1), pp. 61-84.

Ezgi, S. (2021). Voluntary childlessness: A review of the factors underlying the decision not to have children. *Studies in Psychology*, **41**(1), pp. 83-109.

Gillespie, R. (2003). Childfree and feminine: Understanding the gender identity of voluntarily childless women. *Gender & Society*, **17**(1), pp. 122-136.

Griffiths, T. L. (2004). Finding Scientific Topics. *PNAS*.

Hintz, E. A., Haywood, A. (2021). Media frames of voluntary childlessness in the United States from 1989 to 2018. *Sex Roles*, 84(11), pp. 747-764.

Hintz, E. A., Brown, C. L. (2019). Childfree by choice: Stigma in medical consultations for voluntary sterilization. *Women's Reproductive Health*, **6**(1), pp. 62-75.

Misuraca, M., Spano, M. (2020). Unsupervised analytic strategies to explore large document collections. In *Text Analytics: Advances and Challenge*s. Springer International Publishing, pp 17-28

Moore, J. (2017). Facets of agency in stories of transforming from childless by choice to mother. *Journal of Marriage and Family*, **79**(4), pp. 1144-1159.

Morison, T., Macleod, C., Lynch, I., Mijas, M., Shivakumar, S. T. (2016). Stigma resistance in online childfree communities: The limitations of choice rhetoric. *Psychology of Women Quarterly*, **40**(2), pp. 184-198.

Neal, J. W., Neal, Z. P. (2021). Who are the Childfree?. *PsyArXiv* Preprint, 9.

Palmer, E. (2019). The "empty womb" in the therapy room? The taboo and potency of the other than mother/childfree body. *Psychotherapy and Politics International*, **17**(3).

Park, K. (2002). Stigma management among the voluntarily childless. *Sociological Perspectives*, **45**(1), pp. 21-45.

Rivera, E., Logan, E. S., Price, A. (2024). The "childfree" movement: How individuals negotiate identities on reddit. *The Journal of Public and Professional Sociology*, **16**(1).

Röder, M., Both, A., Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399-408.

Zambon, I., Rontos, K., Reynaud, C., Salvati, L. (2020). Toward an unwanted dividend? Fertility decline and the north–south divide in Italy, 1952–2018. *Quality & Quantity*, 54, pp. 169-187.

doi.org/10.26398/asaproc.00138

Mass tourism and local housing: An exploratory textual analysis

Agostino Stavolo[b], Rosanna Cataldo[a], Maria Gabriella Grassia[a], Marina Marino[a]
[a] Department of Social Sciences, University of Naples Federico II, Naples, Italy.
[b] Faculty of Education, Free University of Bozen, Bozen, Italy

1. Introduction

Tourism is one of the most dynamic and promising industries globally, playing a vital role in the economic growth and cultural exchange of many regions. Numerous studies have highlighted the positive impact of tourism development on the receiving areas (Mikulić, 2021; Mathieson and Wall, 1982), including the preservation of the natural environment through ecotourism initiatives, the revitalization of local communities by fostering cultural exchange, and the stimulation of economic growth through job creation and increased revenue. Tourism can lead to the conservation of heritage sites, promotion of local crafts and traditions, and improvement in infrastructure and public services, benefiting both tourists and residents (Dłużewska and Giampiccoli, 2021). However, tourism also brings significant challenges and negative impacts that need to be addressed. One major issue is environmental degradation, which includes increased pollution, loss of biodiversity, and strain on natural resources. The influx of tourists often leads to higher volumes of waste, greater energy consumption, and water shortages, particularly in areas with limited resources. Additionally, the construction of tourism infrastructure can lead to habitat destruction and fragmentation, threatening local wildlife and ecosystems (Kamrowski et al., 2021; Podhorodecka and Dudek, 2019). Destinations worldwide are becoming aware of the fragility and unsustainability of overtourism that dominated their growth models (Mikulić et al., 2020).

Tourism can lead to cultural disruption and social issues within local communities, and it can create social tensions, particularly in destinations where the local population feels overshadowed or marginalized by the tourism industry. This phenomenon is often referred to as 'tourism gentrification', where traditional lifestyles and community structures are displaced by tourism-driven development (Lopes et al., 2019). Economically, while tourism can be a significant source of income, it can also lead to economic dependence on a single industry, making regions vulnerable to fluctuations in tourist numbers due to seasonal changes or economic downturns. Additionally, the benefits of tourism are often unevenly distributed, with profits frequently flowing to external investors rather than the local population (Anderson et al., 2015). The benefits of tourism are often unevenly distributed, with profits frequently flowing to external investors rather than the local population (Anderson et al., 2015). The proliferation of lodging facilities, particularly private rental accommodations such as tourist apartments and bed and breakfasts, plays a significant role in the displacement of residents. As demand for these private rentals increases, property prices and rents escalate, making it increasingly difficult for locals to afford housing. This housing crisis can force residents into precarious living conditions or even drive them out of their communities altogether. The quality of life for residents in tourist-heavy areas can also deteriorate due to overcrowding and increased traffic. The influx of visitors can overwhelm local infrastructure, leading to congestion, noise pollution, and wear and tear on public facilities. However, to harness its potential effectively, it is crucial to plan and develop tourism purposefully and sustainably, balancing the environmental, economic, and social aspects of society to ensure long-term benefits and minimize adverse effects.

In this context, we aim to understand how these dynamics are perceived and addressed by organizations working at the intersection of mass tourism and housing challenges. We analyse the Facebook pages of several national organizations and associations engaged in communication and advocacy on these issues. By applying latent Dirichlet allocation (LDA), we identify the key themes

and concerns related to mass tourism and housing rights. This analysis sheds light on the issues raised by these associations and provides actionable insights to enhance their advocacy efforts and propose sustainable improvements.

2. Background

The proliferation of short-term rentals (STR) is a phenomenon that affects all European cities, particularly historic centres and the most touristically attractive areas. In 2023, Italy recorded over 134 million arrivals and 451 million overnight stays in accommodation facilities, marking the highest levels ever observed in tourism statistics. Within this context, STRs constitute an essential part of the accommodation supply[1]. At the European level, in 2023, guests spent approximately 719 million nights in private rental accommodations booked through online platforms such as Airbnb, Booking, Expedia Group, and TripAdvisor, representing an increase of 20.5% compared to 2022[2]. Specifically, with nearly 600,000 accommodations, Italy ranks second in the world after France for STR offerings on the Airbnb platform, which today represents the largest marketplace within the sharing economy or platform capitalism (Srnicek, 2016; Celata, 2018). More than 608,000 accommodations are available on the Airbnb platform in Italy in 2023, with supply concentrated mainly in three regions: Tuscany (12.9% of the total), Sicily (11.4%) and Lombardy (11.1%)[3].

The primary losers in the Airbnb system are the areas where the platform is most pervasive, specifically cities and their most tourist-attractive zones. While Airbnb can provide some economic benefits, these are often outweighed by negative environmental and socio-economic impacts (D'Eramo, 2022; Gainsforth, 2019). Specifically, there is an increase in real estate prices due to the conversion of housing units into short-term rentals, reducing the availability of affordable housing for residents. Additionally, there is a rise in professionally managed tourist apartments, often handled by intermediaries who take a significant portion of the profits, which primarily benefits large property owners and real estate companies. Furthermore, the phenomenon of touristification leads to significant environmental impacts, such as increased energy consumption, traffic, emissions, and waste. It also causes socio-economic issues, including gentrification, socio-spatial exclusion, social desertification, and the impoverishment of the identity and functionality of urban areas. In summary, while the Airbnb model can support the local economy, it also heavily impacts the areas where it is established, impoverishing them, and deepening social inequalities and territorial disparities (Iovino, 2021).

3. Methodology

The aim of the work is to define the main issues highlighted by national associations regarding the problem of mass tourism and which are the critical issues concerning the right to housing. To comprehend the impact of mass tourism on housing rights, we employed a methodology focused on analysing social media content. Specifically, we targeted several national organizations and associations dedicated to mitigating the adverse effects of mass tourism, specifically to housing rights and preservation of cultural heritage. We chose five specific Facebook pages for this preliminary study:

- Set Napoli - Campagna Resta Abitante: This page represents the city assembly of the Neapolitan node of the Set Network and the Campaign Stay Inhabitant, which focuses on rights in the context of tourism (4,279 followers).

[1] https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,Z0700SER,1.0/SER_TOURISM/DCSC_TUR_OCCYEAR/IT1,122_54_DF_DCSC_TUR_7,1.0

[2] https://ec.europa.eu/eurostat/web/experimental-statistics/collaborative-economy-platforms

[3] https://www.ansa.it/canale_viaggi/notizie/news/2024/03/01/in-italia-608mila-airbnb-il-35-in-toscana-sicilia-e-lombardia_1ef4b3be-c5b8-4fc5-a71b-969dc1a09110.html#:~:text=I%20608mila%20alloggi%20Airbnb%20disponibili,le%20varie%20citt%C3%A0%20e%20territori.

- Pensare Urbano: Established in autumn 2018, this initiative aims to understand and address the housing emergency in Bologna (2,780 followers).
- Mi Riconosci?: A national movement and association founded in 2015 that seeks to obtain greater dignity for cultural work and advocate for a structural reform of the Italian cultural system, benefiting both workers and citizens (47,506 followers).
- Italia Nostra: A national association dedicated to protecting the nation's historical, artistic, and natural heritage (21,769 followers).
- SET - Sud Europa di fronte alla turistificazione: This group consists of movements resisting the processes of touristification affecting cities (2,126 followers).

We collected 2,373 posts from their Facebook pages. Figure 1 shows the distribution of the number of posts by each Facebook page. Notably, "Italia Nostra" and "Mi Riconosci?" account for more than half of the total posts.

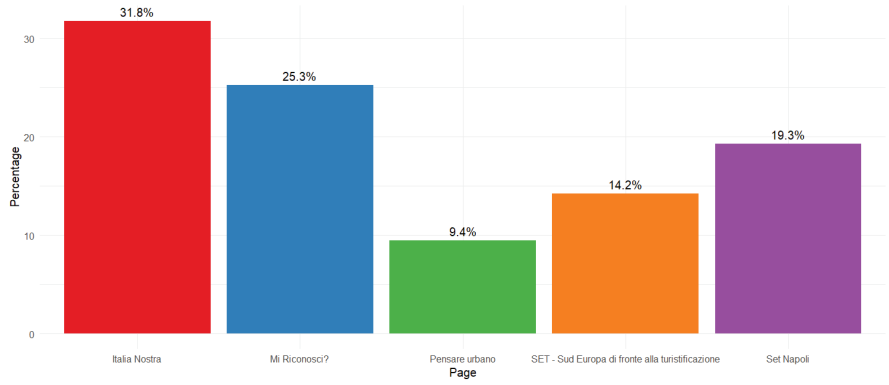


Figure 1 – Distribution of posts by Facebook page

Once the documents were atomized into basic components, preprocessing steps were necessary to reduce linguistic variability. To ensure consistency among the various linguistic variants, additional normalization procedures were implemented, such as correcting misspelled terms, modifying word accents, eliminating numerical characters, and standardizing text by converting characters to lowercase (Misuraca and Spano, 2020). The grammar tagging process was considered, which assigns each term its respective part of speech (POS), identifying nouns, adjectives, and verbs for vocabulary construction. Subsequently, lemmatization was performed, where each term was reduced to its canonical form: verbs were conjugated to the present infinitive, while nouns and adjectives were converted to the masculine singular form. Once the documents were pre-processed, the vocabulary was constructed, recording the number of occurrences of each term (type) in the document collection. Stop-words (such as prepositions, conjunctions, adverbs, etc.) were removed from the vocabulary list, and occurrence trimming was performed, eliminating all terms that appeared with a frequency of less than three. At the end of the process, we have 2373 documents, 24.157 tokens and 4.258 types. So, we apply a topic modeling technique, such as latent Dirichlet allocation - LDA. LDA is a generative probabilistic model of a corpus and is used in unsupervised machine learning algorithms (Blei et al., 2003). The basic idea is that documents are represented as random mixtures of latent topics, where each topic is characterized by a distribution of words. The Dirichlet distribution is used to determine the distribution of topics for each document. In the generative process, the results of the Dirichlet distribution are used to assign words in a document to various topics. According to the model, documents are directly observable, while topics, the distribution of topics per document, and the classification of each word on the topics per document are hidden structures. By employing LDA, we analysed the texts to uncover the main themes and issues discussed concerning mass tourism and housing rights (Istat, 2023). LDA demonstrated its superior performance in specific contexts (Kalepalli, 2020; Bellaour, 2021).

4. Preliminary results

To determine the optimal number of topics, we used the measures proposed by Deveaud et al. (2014) and Griffiths et al. (2004). Devead (2014) proposed a metric that employs the Jensen-Shannon divergence to determine that value of topics that maximizes the distance between each pair of topics; Griffiths et al. (2004) employ a Bayesian method based on the computation, through Gibbs sampling, of the posterior probability of a set of models according to all the words of the corpus, namely the likelihood *Pr (w | K)*. The higher the value, the greater the likelihood of the number of topics. Figure 2 shows the values of these measures and based on the convergence point of the two metrics, we identified *k=6* as the optimal number of topics.

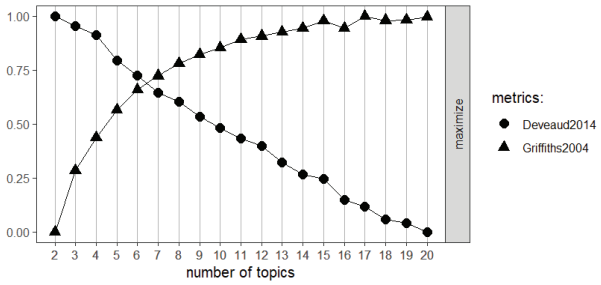


Figure 2 – Selection of topics

In Figure 3, we list ten terms associated with each of the identified topics, categorized with the following labels. The assignment of labels to each topic involves analysing the words most strongly connected with it (Kennedy et al., 2021) and we read the posts related to each topic. Subsequently, we refer to the topic using the designated label. The topic of housing emerges as a critical issue, particularly in cities with high student populations, such as Bologna, where increasing tourist demand has driven up rental prices. The analysis highlights the need for policies to regulate rents, ensure affordable housing, and protect tenants' rights. These measures aim to address broader urban challenges like gentrification and the displacement of local residents, ensuring equitable access to housing for all. In the context of associations, the study explores the role of these organizations in managing environmental and territorial policies at regional and national levels. Efforts are directed towards conserving green spaces, reducing carbon emissions, and managing waste effectively. Environmental education is also emphasized as a way to align tourism development with sustainability goals, ensuring a balance between economic growth and ecological preservation. The research also sheds light on labour relations within the public sector, particularly concerning contracts in the cultural heritage field. The focus is on improving working conditions by advocating for fair wages, job security, and employee benefits, as explained by the topic called "contract". Workers' unions play a pivotal role in these negotiations, ensuring a motivated and stable workforce capable of effectively managing the pressures associated with tourism in heritage-rich cities. Also, the management of cultural heritage is further explored through national projects aimed at preserving historical sites and fostering cultural activities. The study highlights efforts to engage communities in the conservation and promotion of local traditions, often through initiatives like monument restoration and the creation of cultural institutions. These projects are positioned as essential for preserving the historical and cultural identity of communities while fostering broader social engagement. Urban development is another topic of interest, with an emphasis on creating public spaces that support both residents and tourists. The study emphasizes the importance of infrastructure that promotes social inclusion and community participation. Public parks, community centres, and other shared spaces are seen as vital to fostering social interaction, enhancing the quality of urban life, and ensuring a cohesive community. Lastly, the museum theme is central to the sustainable management of cultural heritage. The study explores strategies for promoting sustainable land use, reducing pollution, and implementing ecological practices in urban settings.

Balancing tourism with cultural conservation remains a central challenge, and the study underscores the importance of public engagement in cultural activities. By fostering such participation, policymakers aim to create liveable cities that harmonize economic, cultural, and environmental goals.

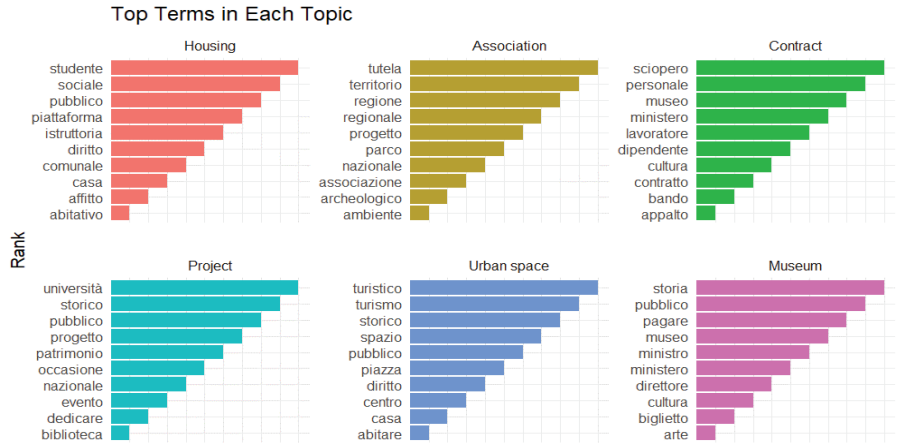


Figure 3 – Top ten terms associated to topics

5. Conclusions

Our analysis of social media content from various national organizations and associations reveals insights into the multifaceted impacts of mass tourism on housing rights and cultural heritage. The study highlights several critical issues raised by these associations, focusing on the economic, environmental, and socio-cultural dimensions (Lopes et al., 2019; Anderson et al., 2015). These organizations actively work to counteract the rise in rental prices, real estate speculation, and the eviction of local residents—phenomena exacerbated by mass tourism and urban transformations tailored to tourists. Additionally, they emphasize the importance of safeguarding the rights and decent wages of workers while ensuring restricted and equitable accessibility to cultural facilities. While mass tourism can drive economic growth and cultural exchange, it also presents significant challenges that require immediate attention through sustainable policy interventions.

Our analysis, utilizing latent Dirichlet allocation, provides evidence-based insights into these challenges. These insights can inform targeted recommendations for policymakers and stakeholders, addressing key issues such as housing affordability, workers' rights, and community well-being. By recognizing and mitigating these impacts, it is possible to enhance the living conditions of affected communities and ensure the sustainable development of tourist destinations.

References

Anderson, L. G., Rocliffe, S., Haddaway, N. R., Dunn, A. M. (2015). The role of tourism and recreation in the spread of non-native species: a systematic review and meta-analysis. *PloS One*, **10**(10).

Bellaouar, S., Bellaouar, M. M., Ghada, I. E. (2021). Topic modeling: Comparison of LSA and LDA on scientific publications. In *Proceedings of the 2021 4th International Conference on Data Storage and Data Engineering*, pp. 59-64.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, pp. 993-1022.

Celata, F., Hendrickson, C. Y., Sanna, V. S. (2017). The sharing economy as community marketplace? Trust, reciprocity and belonging in peer-to-peer accommodation platforms. *Cambridge Journal of Regions, Economy and Society*, **10**(2), pp. 349-363.

D'Eramo, M. (2022). *Il selfie del mondo: indagine sull'età del turismo da Mark Twain al Covid-19*. Feltrinelli Editore.

Deveaud, R., SanJuan, E., Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, **17**(1), pp. 61-84.

Dłużewska, A., Giampiccoli, A. (2021). Enhancing island tourism's local benefits: A proposed community-based tourism-oriented general model. *Sustainable Development*, **29**(1), pp. 272-283.

Gainsforth, S. (2019). *Airbnb città merce: Storie di resistenza alla gentrificazione digitale.* DeriveApprodi.

Griffiths, T. L., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, **101**, pp. 5228-5235.

Iovino, G. (2021). Affitti turistici a breve termine nelle città d'arte. La "airificazione" di Napoli e i suoi impatti. *Bollettino dell'Associazione Italiana di Cartografia*.

Istat. (2023). *Il turismo culturale in Italia*. Istituto Nazionale di Statistica. Retrieved from https://www.istat.it/it/files/2023/10/Il-turismo-culturale-in-Italia_Ebook.pdf

Kalepalli, Y., Tasneem, S., Teja, P. D. P., Manne, S. (2020). Effective comparison of LDA with LSA for topic modelling. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS),* pp. 1245-1250.

Kamrowski, R. L., Limpus, C., Moloney, J., Hamann, M. (2012). Coastal light pollution and marine turtles: Assessing the magnitude of the problem. *Endangered Species Research*, **19**(1), pp. 85-98.

Kennedy, I., Hess, C., Paullada, A., Chasins, S. (2021). Racialized discourse in Seattle rental ad texts. *Social Forces*, **99**(4), pp. 1432-1456.

Lopes, A. C., Rodrigues, E. B., Vera-Cruz, R. (2019). Tourism gentrification. In *IOP Conference Series: Materials Science and Engineering,* IOP Publishing, **471**(9).

Mathieson, A., Wall, G. (1982). *Tourism, Economic, Physical and Social Impacts*. Longman, Harlow, United Kingdom

Mikulić, J., Vizek, M., Stojčić, N., Payne, J. E., Časni, A. Č., Barbić, T. (2021). The effect of tourism activity on housing affordability. *Annals of Tourism Research,* **90**.

Misuraca, M., Spano, M. (2020). Unsupervised analytic strategies to explore large document collections. *Text Analytics: Advances and Challenges*. Springer International Publishing, pp 17-28

Podhorodecka, K., Dudek, A. (2019). Disadvantages connected with the development of tourism in the contemporary world and the concept of sustainable tourism. *Problemy Ekorozwoju*, **14**(2), pp. 45-55.

Srnicek, N. (2017). *Platform Capitalism.* Polity Press, Cambridge Malden, MA.

doi.org/10.26398/asaproc.00139

Multiverse analysis: Heuristic for multi-model inference or theory of the model error?

Venera Tomaselli, Giulio Giacomo Cantone
Department of Economics and Business, University of Catania, Catania, Italy.

1. Introduction

The typical procedure for estimation of unknown quantities involves identifying the mathematical relations between such *estimand* and a sample from known variables. The estimate is often tuned by finding a value that minimise the errors between the predicted estimates and the known values. Such procedures are the 'statistical models', which also can associate the estimate with auxiliary inferential statistics, such as the standard error of the estimate or the *p*-values.

The credibility of a model also depends on the possibility to replicate the results of the model on new samples. Nevertheless, consider such paradox: if one had a larger sample of known values available, one would ideally prefer to include them in the estimation process rather than using them to test the model's performance. This statement is particularly true for small samples, or for samples with many relevant categorical variables influencing the variation in the predicted values. For this reason, when the sample is small or it does not adequately represent the phenomenon under consideration it also makes sense to verify several alternative versions of the model in parallel (Burnham and Anderson 2002), whether simpler or more complex. If the inferences are mutually consistent, it can be stated that the underlying hypothesis is "conceptually replicated", meaning it is robust with respect to changes in the underlying assumptions of the methods of inference.

Nevertheless, testing more alternative models ('specifications') on the same sample, especially if tightly related, implies to run very similar tests, hence inflating the probability to observe lower *p*-values of them. This issue is known in literature as the family wise error rate or FWER (Tukey 1991). It is indeed a malicious practice to slightly alter a model until a significant $p < \alpha$ occurs, possibly randomly, then hiding the number of attempts before reaching such a desired outcome ('*p*-haking', 'data dredging', see Stefan and Schönbrodt, 2023).

Multiverse analysis (MA) is a practice for testing alternative models with a methodical, highly structured, scheme, such that it would be not properly possible to hide undesired specifications (Steegen et al. 2016; Semken and Rossell, 2022; Cantone and Tomaselli, 2024b). In synthesis, it can be summarised with five steps:

1. Identification of the generic model for testing the hypothesis.
2. Identification all the decisions involved for such model, or 'meta-modeling'.
   An example of a decision is: "what should be the link function of this model?"
3. For every decision, establish a finite set of alternatives.
   An example of alternative decisions: *logit* vs. *probit*. as link functions.
4. Estimate parameters for all the combinations of models (specifications). The set of alternatives can be called the 'pseudo-priors' (prior) of the multiverse.
   In the aforementioned example, with 2 binary decisions, the set of combined specifications equals fitting 4 models. Typically, MA run thousands of specifications.
5. Get a 'pseudo-posterior' (posterior) multiverse distribution of estimates, plus a distribution of p-values (i.e. a P-curve, see Simonsohn, Nelson, Simmons, 2014).

Such methodology has been pioneered by Patel, Burford, and Ioannidis (2015), who noticed a concerning pattern in the estimation of the effects of drugs on mortality: due to the aforementioned 'technical deflation' of *p*-values through switching models, they noticed a statistical paradox: two different models for the same *estimand* can show two significant effect, of opposite sign. They

branded such paradox a "Janus effect"; and it can be noticed that it counts as the extreme case of what Gelman and Tuerlinckx (2000) characterised as "type sign error". Janus effect represents a limit in model averaging: should one be allowed to average estimates in presence of such a illogical bivariate distribution? Assessing when two options are truly equivalent alternatives is largely arbitrary and depends on traditional methodological knowledge about "what works best" more than on clear parametric offsets. In this regard, as showed by Del Giudice and Gangestad (2021), an approach to meta-modelisation without a proper assessment of the admissible causal inferences could lead into severely misguided decisions. In this study, this limitation is subjected to formal analysis for the development of a theory of MA.

2. The Cantone-Tomaselli notation

MA involves the comparison of empirical estimates derived from a systematically differentiated multiplicity of specifications of the same conceptual statistical model. The difference between the generic multimodel approach and the multiversal inference lies in how the specifications are elicited. Multiversal methods proceed not by justifying a finite set of alternative specifications based on a theory or a tradition, but, more in detail, by providing set of elements of differentiation for analytical decisions in the phases of specification of the model. The combinations of these elements then form the set of alternative specifications.

This set can be represented as a list of symbolic values for the $\boldsymbol{q}$-`state' of a decision:

$$Q: \{q_1, q_2, \dots, q, \dots, q_v\}$$

Here $\boldsymbol{q}$ is a finite vector of q-states and v is the number of decision involved in the meta-model (or "multiversal model", see Cantone and Tomaselli 2024a).

For example q_1 can represent the question: "should the variable Z by included as a control? In this case $q_1 \coloneqq 1$ would mean "yes", and $q_1 \coloneqq 0$ means "no". Then q_2 can represent the question: which method should be implemented to fill the missing values? $q_2 \coloneqq 0$ would imply "no imputation for the missing values", while $q_2 \coloneqq A$ would indicate "Method A", then $q_2 \coloneqq B$ for "Method B", etc. The setting operator ":=" gives the insight for the metaphor of a state of 'switches' for the $\boldsymbol{q}$ vector: if in a generic multimodel inference, the analyst provides a list of different blueprints for an inferential engine, with a multiversal model Q the analyst provides a list of all the admissible states of the inferential engine. Each unique combination of the q-states of $\boldsymbol{q}$ forms a specification j of the multiversal model.

The compact method of Cantone and Tomaselli (2024a) for representing the specifications as strings of symbols has an immediate benefit: symbols are relatively customary but the symbol '0' holds a special meaning, since it indicates the concept of `absent state' (the switch is turned off) of a decision, which the authors interpret as a simpler state of the model. The presence of a 0 in the string does not imply that the decision involved is binary. In the provided examples, absence of imputation is still symbolised as a 0, since it is the simplest (least demanding) solution, but are presented multiple alternatives for imputation. It follows that the number of $q \neq 0$ symbols in a string counts as the degree of complexity of that specification. In addition, it can be noticed that the Hamming distance (Hamming d), that is the number of not identical symbols in the same position between two vectors, counts as a numeric distance between 'priors', which is a remarkable property for sensitivity analysis (Cantone and Tomaselli, 2024b).

3. Formal characterisation for multiverse analysis

With this formalism is possible to introduce an axiomatic characterisation of model averaging through a multiversal procedure. Consider the generic θ numeric estimand and its estimate $\hat{\theta}$. Canonical sampling theory assumes that a correctly specified estimator would generate a distribution

$$\widehat{\Theta}: \{\widehat{\theta_1}, \dots, \widehat{\theta_k}\}$$

such that the distribution of the sampling errors

$$\theta - \hat{\theta}$$

depends on the proprieties of the k-samples. From the canonical sampling theory are derived the concepts of parametric standard error and intervals of confidence.

Young and Holsteen (2017) derive the concept of model error through an analogy with the canonical sampling theory. Instead of assuming the irrelevance of the specification a priori, one can assume the sample as adequately powerful and the sampling error as trivial and irrelevant.

The analogy of Young and Holsteen is generally not sufficient to justify the approximation of the identification of the *estimand* with the location of its estimates across the multiple models: in canonical sampling theory, this approximation is guaranteed for unbiased estimators when the probabilities of the events of units to be sampled are independent and identically distributed (IID). This feature holds for multiple samples and re-sampling techniques (bootstrap, etc.). Indeed, some applications of multimodel methods (e.g. model averaging) assume that the probability of inclusion of the specifications can be treated also as independent statistical events, even if this clearly violates the aforementioned tenet of the FWER. In practical terms: if the analyst can theorise that the difference in estimation is switching from *logit* to *probit* as someway relevant, it is rather intuitive that this difference must be inferior to the difference between switching the inclusion of a correlated control variable. In other words, not all pairs of 'multimodels' differ in the same way, and not of sets of models have an uniform variation in their 'priors': one can draw a lot of models with minimal differences and brand it a 'multimodel' inference, but the variation of the estimates would not be representative of the uncertainty due to the degree of freedom in the methodological flexibility of the inference mentioned by Young and Holseen (2017).

By the peculiar methods of construction by systemic combinations, multiversal methods reject explicitly assumption of IID, since the inclusion of a specification is co-dependent to the inclusion of all the other specifications with Hamming's *d* equals 1. Even if the probability of inclusion of a specification is still not treated as IID, the sample of specification is treated as if representative of the 'model error' theorised by Young and Holseen (2017). To justify this heuristic, it possible to trace some conjectures. The starting point to define the axiomatisation of the multiversal method is the definition of the concept of 'epistemological equivalence' of two values for a statistic. Epistemological equivalences are defined by the conditions posed by the test they are associated with. For example, in a test, the most typical condition is defined by the α threshold which demarcates the space of possibilities for a *p*-value in two: lesser and greater of it. With this example in mind, a formal definition is provided:

- **Epistemological equivalence**: Within a system that evaluates differences, two quantities are epistemologically equivalent if their difference does not count as sufficient evidence to alter a belief about the states they represent.

Now it is possible to define the first axiom of the multiversal method:

- **First axiom of multiverse analysis**: For any estimation for θ exists a non-empty interval $\Theta_{\tilde{\epsilon}}$ defined by its extremes $(\tilde{\epsilon}_+, \tilde{\epsilon}_-)$, and it holds
$$\Theta_{\tilde{\epsilon}}: \{\theta + \tilde{\epsilon}_+ > \theta_{\tilde{\epsilon}} > \theta + \theta + \tilde{\epsilon}_-\}$$
such that any $\theta_{\tilde{\epsilon}} \in \widetilde{\Theta}_{\tilde{\epsilon}}$ is epistemologically equivalent to θ.

It is followed by a theorem:

- **Theorem**: Since $\theta_{\tilde{\epsilon}}$ exists, then exists at least a non-finite set of model specifications
$$J_\theta: \{j_1, j_2, \dots, j, \dots\}$$
such that for all of their estimates for $\hat{\theta}_j$ holds that
$$\hat{\theta}_j \in \Theta_{\tilde{\epsilon}}$$

Consider also:

- **Second axiom of multiverse analysis:** If J_θ exists then must exist at least a finite set made of k specifications

$$\hat{J}: \{\hat{j}_1, \hat{j}_2, \dots, \hat{j}, \dots, \hat{j}_k\}$$

such that holds

$$\lim_{k\to\infty} Pr.\left(\hat{\theta}_{\hat{j}} \in \Theta_{\tilde{\epsilon}}\right) \to 1$$

The meaning of the second axiom is that exists at least a collection of specifications $\hat{J}$ that, randomly or deterministically, counts as a representative sample of the process generating the all the epistemological equivalent specifications for the test on θ.

Within the implicit theory of the heuristic of MA, this methodological device holds in the form of

- **Multiversal conjecture:** for $k \to \infty$**,** the set Q (that is, the multiversal *prior*) must converge to J_θ.

This is an ideal condition for Q, yet it suffers from some limitations. Indeed, in practice this convergence is almost never true, for two reasons. The size of multiverse is almost always set after a conveniently reasoned justification and does not follow any specific form of power analysis to generate at least a target *k*, so posing a $k \to \infty$ is often an irrelevant or inconvenient assumption. In addition, by its own definition, epistemological equivalence depends on external parameters adopted for decision-making; be it the actual α of the test, or other heuristics, e.g. a threshold for Bayes factor, etc. In practice, as long as a multiverse has a high degree of internal coherence in how its estimates and other *posterior* statistics can be interpreted, it can count as a signal that the resulting multi-model inference corroborates or rejects a hypothesis.

Essentially, the multiversal conjecture regards the capacity of multiversal inference to be more consistently right than singular tests. However, if the weak conjecture is true, the setup of a multiverse sufficiently sized (sufficiently 'powered') can be treated as a valid $\hat{J}$ for the estimation of the uncertainty regarding θ.

4. Pruning the multiverse: A network theory

It should be clear from the axiomatic characterisation of MA that the multiversal methods enforce a system of mutual dependencies in the selection of $\hat{j}$, which contrasts with other practices in multimodel approaches for model averaging. Cantone and Tomaselli (2024a, 2024b), by introducing the Hamming's *d,* characterise this mechanic of jointed inclusion as the growth of a regular network where each specification is a node connected to an equal number of nodes. In modern theory of networks (Crane, 2018), nodes can hold many attributes. So, the node can encapsulate information on *prior* features and *posterior* statistics. This congruence between a string and a network can be demonstrated through the equivalence of the *d* with the length of the shortest path of the nodes representing the two specifications.

In many regards, this characterisation can be considered an expansion of the *p*-grid proposed by Steegen et al (2016). Nevertheless, a generalised characterisation of the multiverse as a network should account for multiverses that are not regular networks, because they are pruned of a component. Pruning happens when two decisions which, taken independently are plausible and allowed in the prior of the multiverse, and yet when occurring simultaneously they would certainly lead into a bias of the estimate. An example is provided: consider a sample stratified across a G categorical variable, and assume that the research notices absence of uniformity in the estimates across the *g* groups from G. Then it could be the case to include an adjustment of $\hat{\theta}$ for the fixed effects of G=*g*. In spite of the decision to include a fixed effect then would preclude other adjustments for any other Z variable, which is collinear in regard to G. For example, typically if one includes a fixed effect for a country variable, then other variables should not be normalised for the population of that country to avoid a bias of multicollinearity; however if the fixed effect is absent then observation should be regularised by the population.

This is a rather intuitive example of the kind of error, which one can fall by just running "all the combinations" envisioned in the meta-model. In reality the kind of "pitfalls" mentioned by Del Giudice and Gangestad regards more the problem of collider bias when Z is correlated with the *explanand* outcome Y yet it is not its antecedent (chronologically or logically). In cases where Y is a cause of Z, and the independent variable X is it too, the estimator of $\hat{\theta}$ must never be adjusted for Z. Yet, since both X and Y would look correlated to Z, one is lead to include in the multiverse the variant with this adjustment. As explained by Giudice and Gangestad, such an inclusion would bias exactly half of the whole and manifest the Janus effect. The decision must be pruned out of the prior.

In cases like these, a multiverse is still a network, yet made of regular clusters: each node within the cluster has the same degree of links. In this case, the string ceases to represent the material operation involved in the composition of the model and it represents the prior knowledge about the effect of the composition of the multiversal model.

Conclusions

The proof of the multiversal conjecture is, without doubt, the most obvious frontier development of multiverse analysis. Realistically, a proof seems unlikely without a rather rigid set of assumptions; however, the regular structure of the multiverse should help clarify that the main obstacle to the statistical representativeness of the multiverse, in terms of a plausible value distribution for the *estimand*, does not appear to be a problem of variance but rather one of introducing bias. In this regard, it is important to note that Cantone and Tomaselli, in response to Del Giudice and Gangestad, argue that in simulation studies on structures of causal collision it is possible for two opposing sources of bias to cancel each other out. In general, the entire theoretical framework advanced by Cantone and Tomaselli relies not so much on the multiversal conjecture itself but on the specific assumption that if the decisions that introduce bias into the multiverse can be treated as statistically independent events, for k tending to infinity, they can be reduced to an inflation of the variance of the distribution.

A remarkable application of MA concerns the estimation in models of measurement (Greco et al., 2019; Leibel and Bornmann, 2024). These are models where the *estimand* is not a coefficient but a numeric feature in itself. The traditional approach for estimation in measurement models is defined as reflexive: a missing X is reconstructed by eliciting a set of Y variables upon which X exerts an effect on. The main problem with these models is in the free parameters: if one does not know the values of the independent *estimand* variable, it is challenging to determine the coefficients of the effect exerted on the known variables. This is usually resolved through a test of unifactorial factor analysis, i.e. testing the hypothesis of the existence of a singular common factor. The estimator derived by the factor analysis reconstructs a normalized estimate of X.

In this reflexive approach, there is relatively little that needs to be opted for, but the situation is quite different in formative approaches, which are very common in applications involving human action. In a formative approach to measurement, it is assumed that many independent variables are known, and a criterion is established by which they contribute to the variable to be measured. The criteria for the formation are numerous and quite heterogeneous, and in practice, each specific model can be reduced to a form in which many X variables are synthesised according to a aggregative function and a weighting scheme, that can be either implicit or explicit.

In the formative approach, the imposition of a multiversal meta-model, connecting the structure of the alternatives criteria as a network of regular clusters allows for a pseudo-posterior distribution of estimates for the variable to measure. Such result enables the verification of whether different criteria lead to a stable and robust ranking or if the measurement of the *estimand* is highly sensitive to small variations in the assumptions of the measurement model's criteria.

References

Leibel, C., Bornmann, L. (2024). Specification uncertainty: What the disruption index tells us about the (hidden) multiverse of bibliometric indicators. *Scientometrics*. https://doi.org/10.1007/s11192-024-05201-1

Burnham, K. P., Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Cantone, G. G., Tomaselli, V. (2024a). Theory and methods of the multiverse: An application for panel-based models. *Quality & Quantity*, *58*(2), pp. 1447–1480. https://doi.org/10.1007/s11135-023-01698-5

Cantone, G. G., Tomaselli, V. (2024b). Characterisation and calibration of multiversal methods. *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-024-00610-9

Crane, H. (2018). *Probabilistic Foundations of Statistical Network Analysis*. Chapman and Hall / CRC.

Del Giudice, M., Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920954925. https://doi.org/10.1177/2515245920954925

Gelman, A., Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), pp. 373–390. https://doi.org/10.1007/s001800000040

Greco, S., Ishizaka, A., Tasiou, M., Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, *141*(1), pp. 61–94. https://doi.org/10.1007/s11205-017-1832-9

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), pp. 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Semken, C., Rossell, D. (2022). Specification analysis for technology use and teenager well-being: Statistical validity and a Bayesian proposal. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *71*(5), pp. 1330–1355. https://doi.org/10.1111/rssc.12578

Simonsohn, U., Nelson, L. D., Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), pp. 534–547. https://doi.org/10.1037/a0033242

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), pp. 702–712. https://doi.org/10.1177/1745691616658637

Stefan, A. M., Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346. https://doi.org/10.1098/rsos.220346

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*(1), pp. 100–116.

Young, C., Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), pp. 3–40. https://doi.org/10.1177/0049124115610347

doi.org/10.26398/asaproc.00140

From crisis to crisis: Mapping material deprivation in the EU through the 2000s

Susanna Traversa[a], Enrico Ivaldi[b]
[a] Italian Center of Excellence on Logistics, Transport and Infrastructures (CIELI), Department of Economics, University of Genoa, Genoa, Italy
[b] Faculty of Communication, IULM University, Milan, Italy

1. Introduction

The early 21st century has been marked by significant socio-economic disruptions and global health crises, which have deepened inequalities within and among European Union (EU) member states. Following the 2009 financial crisis, which originated from the collapse of the U.S. subprime mortgage market, European institutions have increasingly focused on developing policies to foster a fair and sustainable recovery. A central objective of the EU has been to promote economic development across the region, with "social convergence" playing a crucial role in achieving this goal (Moscovici, 2017). This commitment to social convergence is reflected in initiatives and funds such as the Europe 2020 Strategy, which aimed to deliver smart, sustainable, and inclusive growth, the European Social Fund Plus (ESF+) and the more recent National Recovery and Resilience Plans. These latter were implemented in response to the COVID-19 pandemic, which severely disrupted economic activities, led to significant job losses, and exacerbated the existing socio-economic inequalities between and within member states (European Commission, 2010; European Commission, 2020).

One area of focus in addressing social inequality has been material deprivation (MD), which refers to the inability to access essential goods and services, putting individuals at a disadvantage within their community (Townsend, 1979; Mack and Lansley, 1985). Despite growing interest in studying deprivation, it remains difficult to pinpoint its causes due to its complex, multidimensional nature. To better understand MD, researchers and Eurostat have broken it down into subcategories such as household deprivation, economic stress, and enforced lack of durable goods (Guio, 2009). Economic stress refers to severe disadvantage leading to social exclusion, while enforced lack indicates the inability to access certain goods. Given the complexity of these issues, the use of synthetic indicators, which combine different dimensions of material deprivation, has gained support in the literature as a way to simplify analysis and interpretation (Traversa et al., 2024; Bossert et al., 2013; Fura et al., 2017).

Recent studies have introduced innovative methods to track the historical progression of material deprivation, as s persistent lack of resources that leads to social exclusion (Calandrino, 2003). By incorporating a temporal dimension, these studies provide better tools for monitoring changes in deprivation and identifying vulnerable regions, helping policymakers create more effective public policies. Key contributions include the use of the Townsend Index in the UK to support local policies (Norman, 2010), the development of new indices like the Genoa deprivation index in Italy (Landi et al., 2018), and a similar approach applied to Poland (Dudek and Szczesny, 2021). The COVID-19 pandemic has sparked further research into its impact on deprivation levels, highlighting the ongoing relationship between material deprivation, health, and environmental factors. These insights are crucial for informing sustainable recovery and green policies across Europe.

While most studies focus on local or national levels, this research examines broader EU-wide trends over the first two decades of the 21st century. Thanks to Eurostat data on economic stress and the lack of durable goods, the study will create an index to measure material deprivation across

EU member states over time, providing a macroscopic view of whether national actions have improved socio-economic conditions and how these efforts, along with major socio-economic shocks, have impacted the EU as a whole.

2. Materials and methods

This study aims to track material deprivation trends in the European Union from 2005 to 2022 by developing a synthetic index. Using data from the EU-SILC module of the Eurostat database, the study ensures consistent and comparable results across EU member states. Material deprivation became a key focus in EU policy around 2004–2005, with formal indicators introduced in 2009 (Fusco et al., 2013), allowing for a thorough analysis using data from 2005 onward.

The index uses widely recognized indicators from the EU2020 strategy to ensure consistency across member states. It focuses on two key dimensions: "economic stress" (financial difficulties in meeting basic needs) and "enforced lack of durable goods" (inability to afford essential items). The analysis includes 26 EU countries, with Croatia and the UK excluded due to missing data, to maintain accuracy and reliability. Even if excluding these two countries may affect the overall trend of the scores, as it introduces a potential bias in the results, the lack of data for Croatia between 2005–2009 and the incomplete dataset for the United Kingdom for some post-2018 variables increase the risk of errors in the index calculation due to the need to estimate these missing values.

Table 1. Overview of material deprivation variables selected. Source: EU-SILC database, Eurostat.

MD Dimension	Eurostat variables	EU-SILC Cod.
Enforced lack of durable goods	Inability to afford a telephone	Ilc_mddu01
	Inability to own a color TV	Ilc_mddu02
	Inability to afford a washing machine	Ilc_mddu04
	Inability to afford one's car	Ilc_mddu05
Economic stress	Inability to heat adequately one's own home	Ilc_mdes01
	Inability to take one week's holiday a year	Ilc_mdes02
	Inability to afford a protein meal or vegetarian alternative, every two days	Ilc_mdes03
	Inability to cope with unforeseen expenses	Ilc_mdes04
	Inability to meet overdue payments	Ilc_mdes05
	Inability to make ends meet	Ilc_mdes09

To perform hierarchical comparative analysis, many researchers employ the adjusted Mazziotta and Pareto index (AMPI$^{\pm}$), a refined version of the original Mazziotta Pareto index (MPI) (Mazziotta and Pareto, 2012). This method, developed by Mazziotta and Pareto, is particularly favoured for its adaptability and ease of replication, making it well-suited for analysing complex phenomena over time. The AMPI$^{\pm}$ is appreciated for its systematic construction process and its ability to manage normalization challenges through a constrained rescaling approach. By addressing issues such as sensitivity to outliers and difficulties in centring averages, the method proves effective for both longitudinal and spatial comparisons (Mazziotta and Pareto, 2018).

Recent studies have compared AMPI$^{\pm}$ with other methods for assessing well-being and inequality, demonstrating its effectiveness due to its clarity and simplicity. The method normalizes indicators within a specific range and centres them around a reference point, allowing for meaningful comparisons across different periods and geographical areas. Additionally, it incorporates penalties during aggregation to reflect the nature of the phenomenon being measured.

The mathematical synthesis of the AMPI$^{\pm}$ index involves these key features, contributing to its adoption in deprivation studies.

Firstly, the transformation of the original data matrix $X = \{x_{ijt}\}$ into a normalized one R= $\{r_{ijt}\}$ is performed, where r_{ijt} is the normalized value and x_{ijt} is the value of the indicator j^{th} for the unit i^{th} at the time t^{th}, while Min_{xj} and Max_{xj} indicate the "goalposts" for the indicator j.

$$r_{ijt} = \frac{x_{ijt} - Min_{x_j}}{Max_{x_j} - Min_{x_j}} * 60 + 70$$

After the normalization, the second step is dedicated to the definition of the "goalposts".

$$Ref_{ijt} \pm \Delta \; with \; \Delta \frac{Sup_{xj} - Inf_{xj}}{2},$$

Ref_{xj} is the reference value, i.e., the value of the indicator j^{th} in a specific unit i^{th} at a specific time t^{th}, and Inf_{xj} and Sup_{xj} are, respectively, the minimum and the maximum value of the indicator j^{th} in all units and all periods. For this index, the goalposts are calculated by assuming 2005 as a reference (the first year of data collection). In this way, each indicator assumes the value of 100 for the reference unit considered in the time occasion considered in all basic indicators; all the other values of each unit for all the time occasions will be expressed about this value, allowing for comparison in time and space. Using this normalization, the range of the normalized values is [70:130].

The last step consists of the aggregation of the $AMPI^{\pm}$ index:

$$AMPI_i^{\pm} = \mu_{ri} \pm \sigma_{ri} cv_i,$$

The $AMPI^{\pm}$ method evaluates each unit's score by splitting it into two components: the mean level (μ_{ri}) and a penalty ($\sigma_{ri} cv_i$). The penalty addresses variability among indicators relative to the mean, aiming to reward units that, despite having the same average, display more balanced indicator values. In the context of material deprivation (MD), where all composite indices are negative (indicating worse conditions with higher values), the average of standardized indicators is adjusted upwards with a positive penalty. This correction ensures that the index properly reflects improvements in material deprivation.

Due to the subjective choices involved in constructing a synthetic index, it is crucial to test its robustness (OECD, 2008). In this case, the validation process used to assess how sensitive the results and rankings are to changes in the index's structure is made by the operation of an Influence Analysis (IA). By iteratively reconstructing the index while excluding different indicators each time, we can assess the robustness of the index. This involves examining how each component influences the outcome, focusing on the absolute mean deviation of ranks in the new rankings compared to the original (Mazziotta and Pareto, 2018; Tarantola et al., 2000). The results of this robustness evaluation are typically presented as country rankings with associated uncertainty bounds, which indicate the range of possible values for each country. The robustness analysis for the material deprivation index shows a very low average deviation (0.55) from the initial rankings, confirming that the index is robust and reliable for the intended analysis.

3. Results

After calculating the $AMPI^{+}$ scores throughout the period, the rankings for each year were constructed by assigning the first rank to the country with the highest score, corresponding to higher levels of social exclusion (Table 2).

Table 2. Extract of the results and rankings obtained following the application of the index to EU26 countries.

EU26 MS	2005		2008		2013		2018		2022	
	$AMPI^{+}$	Rank	$AMPI^{+}$	Rank	$AMPI^{+}$	Rank	$AMPI^{+}$	Rank	$AMPI^{+}$	Rank
Austria	89	24	91	18	90	23	76	24	89	19
Belgium	92	19	91	19	93	19	92	15	91	15
Bulgaria	144	1	125	1	126	1	167	2	104	3
Cyprus	102	10	104	6	111	6	133	4	96	4
Czech Republic	97	14	93	15	97	16	79	21	86	25
Denmark	91	21	89	23	92	20	81	19	88	20
Estonia	101	11	93	16	97	15	92	16	90	17
Finland	91	22	89	24	91	22	81	18	88	21
France	91	23	90	21	90	24	86	17	92	13
Germany	91	20	90	20	91	21	76	22	91	14

Greece	99	13	101	10	113	5	169	1	112	1
Hungary	107	7	105	5	116	3	119	7	95	5
Italy	99	12	102	9	105	8	105	10	92	12
Latvia	121	2	105	4	116	4	123	5	94	7
Lithuania	113	5	98	11	102	11	120	6	93	10
Luxembourg	92	18	90	22	93	18	70	26	87	24
Malta	103	9	98	12	103	9	81	20	90	18
Netherland	88	25	86	26	89	25	72	25	86	26
Poland	115	4	103	7	99	14	95	13	90	16
Portugal	104	8	107	3	105	7	107	8	93	9
Republic of Ireland	94	17	91	17	100	12	94	12	92	11
Romania	121	3	125	2	124	2	152	3	105	2
Slovakia	109	6	102	8	103	10	107	9	94	8
Slovenia	94	16	94	14	107	17	93	14	88	22
Spain	95	15	94	13	122	13	100	11	95	6
Sweden	87	26	87	25	73	26	74	23	87	23

The highest ranking is assigned concerning greater levels of deprivation (high $AMPI^+$ score).

With the aim of better understanding which areas of the European Union have suffered the most the effects of the crisis, a Ward cluster analysis (Ward, 1967) was conducted on the $AMPI^+$ scores calculated for 2005. Using the elbow method, the optimal partitioning was identified into 4 clusters and, subsequently, their intervals of them were calculated. This step is crucial for defining deprivation classes, which can be used to support the study of how deprivation has evolved. With these deprivation classes, it is possible to compare and track changes in deprivation levels from 2005 onward, allowing for a more detailed and informed analysis of socio-economic trends.

4. Discussion

Reconstructing the evolution of the phenomenon over 18 years reveals a general improvement in the socio-economic conditions of the European population. Specifically, the average EU26 score on the $AMPI^+$ index decreased from 101 in 2005 to 93 in 2022. This decline indicates an overall positive trend in socio-economic well-being across the EU during this period.

In 2005, a higher concentration of medium-to-high levels of deprivation was observed in Eastern European and Balkan countries, with Bulgaria in particular exhibiting a notably high score of 144, placing it singularly in the first cluster. This pattern of inequalities at the beginning of the 2000s corroborates existing literature, which indicates that the new EU member states, referred to as the EU-10, display heightened susceptibility to socio-economic disparities. Tracking the evolution of inequalities in Bulgaria, the country has alternated between first and second place in the annual rankings, dropping to third only in 2022. After 18 years of a positive trend, despite a slight increase in 2013, the index score dropped to 104, a decrease of 40 points from 2005.

By 2008, while some eastern European countries like Poland, the Czech Republic, and Slovakia showed improvements, others, such as Romania and Portugal, experienced general worsening conditions due to the initial phase of the financial crisis. This shift led to higher $AMPI^+$ scores (Romania 125 from 120, and Portugal 107 from 104) and a change in their deprivation rankings in the top half of the ranks.

By 2013, the deprivation distribution faced a shift from a vertical to a horizontal divide, less marked, highlighting greater difficulties in southern European countries compared to Scandinavia and Central Europe, which showed more stability despite some fluctuations.

In this phase, Mediterranean countries Mediterranean countries have experienced a significant increase in economic difficulties, with Italy, Malta, and Greece showing worsening deprivation indicators. Specifically, Greece has seen a notable rise in its $AMPI^+$ score, which has reached 114, marking a 14-point increase since 2007. This reflects a substantial rise in economic stress and social exclusion within the country. On the other hand, Italy has reported a sharp increase in the percentage of individuals unable to handle unexpected expenses, which rose from 57.3% in 2005 to 74.9% in 2013. Italy's $AMPI^+$ score also increased, reaching 105, with a 5-point rise since 2007. These data

points underscore the intensifying economic challenges faced by Greece and Italy, exacerbated by the 2008 global financial crisis and the Eurozone debt crisis, along with the subsequent austerity measures that heavily impacted these countries.

The long-term trend up to 2018 indicates positive progress in social inclusion across Europe, largely due to national and EU-level social policies such as the Europe 2020 Agenda. This progress is evidenced by a reduction in the most deprived clusters and a general shift towards lower levels of social exclusion, reflecting advancements in addressing material deprivation. By 2018, the first cluster of high deprivation had been completely emptied, with only the second cluster remaining and now including only two countries: Bulgaria and Greece. By 2022, the most deprived countries were primarily from Eastern Europe, with Greece, Portugal, and Spain also among the top ten most deprived.

The case of Greece, in particular, stands out as significant. By the end of the 18 years, Greece, along with Romania and Bulgaria, occupied the second deprivation cluster, ranked first, second, and third in this cluster, respectively, with $AMPI^{+}$ scores of 112, 105, and 104. Greece's position underscores the persistent challenges it faces in tackling material deprivation and social exclusion, despite broader European progress. The country experienced a 14-point increase in its $AMPI^{+}$ score since 2005, reaching a peak of 117 in 2016. During that year, the most critical dimensions for Greece were related to economic stress, with an average of 38% of the population experiencing deprivation in this area.

Poland presents a notable success story. Its $AMPI^{+}$ score significantly improved from 114.5 in 2005 to 90.2 in 2022, and a change in its ranking from 4^{th} to 16^{th} rank. This positive development is largely attributed to Poland's transition to a capitalist economy and the implementation of effective social reforms, such as the "Family 500+" program. This initiative played a crucial role in reducing poverty by 40% between 2015 and 2018 (World Bank, 2017), showcasing Poland's successful efforts in mitigating material deprivation and enhancing social welfare.

5. Conclusions

The development of the European material deprivation index from 2005 to 2022 marks a significant advancement in our understanding and response to material deprivation across Europe. This index offers a continuous view of deprivation trends relative to a baseline year, allowing policymakers to monitor changes over time and evaluate the impact of various interventions. Its longitudinal approach is instrumental in guiding recovery policies and fostering social convergence across the region. The index's capacity to reflect nuanced variations in material deprivation provides valuable insights that can inform targeted policy measures aimed at promoting inclusive growth and equitable development.

Despite its strengths, the study has limitations. Notably, Croatia is excluded from the analysis, despite being an EU member since 2013 it might exhibit similar deprivation sensitivities as other EU-10 countries. Future research could benefit from incorporating additional variables of material deprivation, such as the possession of personal computers or internet connectivity. These factors have gained prominence, especially during the pandemic, when maintaining social connections became increasingly dependent on digital tools. Expanding the index to include such variables could enhance its relevance and accuracy in capturing current deprivation dynamics.

References

Bossert, W., Ceriani, L., Chakravarty, S.R., D'Ambrosio, C. (2013). *Intertemporal Material Deprivation. Poverty and Social Exclusion: New Methods of Analysis.* Routledge: London, (UK), p. 128. ISBN 9781138241343

Calandrino, M. (2003). *Low-Income and Deprivation in British Families. An Exploratory Analysis of the 'Consistent Poverty' Approach to Poverty Measurement Using Data for Great Britain*

Drawn from the Families and Children Study. A Study Carried Out on Behalf of the Department for Work and Pensions. Department for Work and Pensions: London, (UK). ISBN 1843881918.

Dudek, H., Szczesny, W. (2021). Multidimensional material deprivation in Poland: A focus on changes in 2015–2017. *Quality & Quantity*, **55**, pp. 741–763.

European Commission (2010). *EUROPE 2020: A Strategy for Smart, Sustainable and Inclusive Growth*. Publications Office of the European Union: Luxembourg. Retrieved from: eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52010DC2020 (Available online).

European Commission (2020). *Europe's moment: Repair and Prepare for the Next Generation. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions*. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri= CELEX:52020DC 0456&from=EN (Available online).

Fura, B., Wojnar, J., Kasprzyk, B. (2017). Ranking and classification of EU countries regarding their levels of implementation of the Europe 2020 strategy. *Journal of Cleaner Production*, **165**, pp. 968–979. https://doi.org/10.1016/j.jclepro.2017.07.088.

Guio, A.C. (2009). *What Can Be Learned from Deprivation Indicators in Europe?* Indicator subgroup of the Social Protection Committee, 10. Office for Official Publications of the European Communities: Luxembourg. ISBN 978-92-79-12308-5.

Landi, S., Ivaldi, E., Testi, A. (2018). Measuring change over time in socio-economic deprivation and health in an urban context: The case study of Genoa. *Social Indicators Research*, **139**, pp. 745–785. https://doi.org/10.1007/s11205-017-1720-3.

Mack, J., Lansley, S. (1985). *Poor Britain*. London: G. Allen & Unwin. ISBN 0043360831.

Mazziotta, M., Pareto, A. (2012). A non-compensatory approach for the measurement of the quality of life. In *Quality of Life in Italy*; Springer: Cham, Switzerland; pp. 27–40.

Mazziotta, M., Pareto, A. (2018). Measuring well-being over time: The adjusted Mazziotta–Pareto index versus other non-compensatory indices. *Social Indicators Research*, **136**, pp. 967–976.

Moscovici, P. (2017). *Making Economic and Social Convergence Our Political Priority. Closing Remarks, Reinventing Convergence: Towards Economic Resilient Structures.* European Commission: 19 October, Brussel (BE).

Norman, P. (2010). Identifying change over time in small area*s* of socio-economic deprivation. *Applied Spatial Analysis and Policy*, ***3***, pp. 107–138.

OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing: Paris (FR). https://doi.org/10.1787/9789264043466-en.

Tarantola, S., Jesinghaus, J., Puolamaa, M. (2000). Global sensitivity analysis: A quality assurance tool in environmental policy modelling. *Sensitivity Analysis of Model Output* K. M. Hanson, F. M. Hemez*,* eds. Los Alamos National Laboratory, 2005, pp. 385–397.

Townsend, P. (1979). *Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living.* University of California Press. ISBN 9780140221398.

Traversa, S., Ivaldi, E., Forciniti, A. (2024). The European Union and material deprivation: Measuring the evolution of inequalities over the 2000s. *Sustainability*, **16**(14), 6148. https://doi.org/10.3390/su16146148

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), pp. 236–244.

World Bank Group (2017). *Lessons from Poland, Insights for Poland: A Sustainable and Inclusive Transition to High Income Status.* World Bank, Washington, DC. Available online: http://hdl.handle.net/10986/28960 (accessed on 5 July 2024).

doi.org/10.26398/asaproc.00141

The effect of user web journey on browsing behaviour and buying attitude: The case of Sunweb.nl

Furio Urso[a], Nicola Argentino[a], Antonino Abbruzzo[a], Reza Mohammadi[b], Kevin Pak[b], Maria Francesca Cracolici[a]

[a] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy.

[b] Faculty of Economics and Business, Section Business Analytics, Amsterdam Business School,Netherlands

1. Introduction

Digital technologies have significantly transformed the tourism industry by diminishing the role of traditional intermediaries like travel agencies, allowing firms to connect directly with customers (Zhou (2004)). To leverage these direct interactions, it is crucial for businesses to understand online consumer behaviour, obtaining valuable insights into their preferences and buying intent in order to forecast purchases (Huynh et al., 2022, Zhang et al., 2015).

The "customer journey" framework offers a way to map and analyze digital interactions – both observable and unobservable – that reflect consumer interests and actions throughout their e-commerce experience (Dunn and Davis, 2003). This framework helps companies identify customer stages and devise targeted marketing strategies to influence purchasing decisions.

Our study focuses on the pre-purchase phase of this journey, where we analyze website data to understand browsing behaviours and predict purchase intention (Kim et al., 2024, Koehn et al., 2020). In the tourism sector, predicting purchase behaviour is particularly challenging due to the intangible and heterogeneous nature of services (Moeller, 2010). Despite the importance of web data for understanding consumer behaviour, there is limited research on using clickstream data for predicting online tourism purchases (Ypma and Heskes, 2002, Park and Park, 2016).

Previous studies have combined structured and unstructured data for purchase predictions (e.g., Zhu et al.,(2019, Chen et al., 2021), but this study uniquely leverages both traceable and anonymous clickstream data to explore users' browsing behaviour and predict purchases. We aim to address three key questions: Can anonymous user data enhance our understanding of browsing behaviour? Can browsing behaviour be used to identify distinct user profiles? How does shifting between profiles affect purchase predictions?

A two-stage procedure is defined to explore clickstream data from Sunweb.nl that refers to the period from May to August 2022. At first, a mixture hidden Markov model (MHMM) is used to identify browsing profiles by clustering web sequences (sequences of web pages) representing users' activity in the website during each visit. The clustering step considers both anonymous and traceable user data, taking advantage of the whole information of the dataset in order to classify distinct browsing behaviours.

At the second stage, we focus on traceable user data, specifically users who have made multiple visits to the website. A logistic regression model was estimated to explore how visits, classified into different behavioural profiles at the first stage, influence the likelihood of users making a purchase at the end of their web journey. The analysis highlights the existence of specific browsing patterns associated with higher purchase probabilities. By integrating anonymous clickstream data into the clustering process and linking cluster transitions to purchase likelihood, our study gives insights into how past behaviours influence purchasing decisions and bridges theoretical insights with practical applications for improving digital marketing strategies in the tourism industry.

The proposed procedure and methods, i.e. MHMM and logistic regression, are illustrated in Section 2; the results and discussion are presented in Section 3; some final considerations and future developments will follow.

2. Behavioural profiles identification procedure

As aforementioned, the online consumer behaviour has been explored by means a two-stage procedure. At the first step, to identify browsing behaviour profiles, a model-based clustering method such as the mixture hidden Markov model (Vermunt et al. (2008)) is used, accounting for the uncertainty related to Web sequences. These clusters of sequences take into consideration all individual visits (anonymous and tracked users) and are identified with respect to how users move around the website between pages and how they access the website (i.e., access device).

We assume that $M = \{M^1, M^2, \dots, M^K\}$ is a set of hidden Markov models (HMMs) and sequences belonging to the same cluster are generated by the same HMM. MHMM log-likelihood (Van de Pol and Langeheine, 1990) is presented in Equation 1.

$$\ell(\Theta, \omega; y, X) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \omega_{ik} \sum_{u} \pi^k_{u_{i1}} b^k_{u_{i1}}(y_{i1}) \prod_{t=2}^{T} a^k_{u_{i,t-1}u_{it}} b^k_{u_{it}}(y_{it}) \right) \quad (1)$$

where y_i are the observed sequences of length T and each element takes values $card|y_{it}| = R$ that are the web pages, u_i are hidden sequences, for each cluster k: $\Theta = \{\Theta^K\} = \{\pi^k, A^k, B^k\}$ is the set of parameters: initial, transition, and emission probabilities for each sub-model M^k, $\pi^k_{u_1}$ is the initial probability; $a^k_{u_t,u_{t-1}}$ is the transition probability between clicks $t-1$ and t, and $b^k_{u_t}$ is the probability that the hidden state of subject i at time t emits the observed state (a web page). Each sequence y_i is assigned to a browsing behaviour cluster by updating cluster membership $\omega_{ik} = P(M^K|X_i)$, which depends on a set of covariates X_i representing information obtained from clickstream data (e.g. access device).

Then, the number of browsing profiles K is selected by using information criteria such as BIC. The model's parameter estimation is performed through the Estimation-Maximization algorithm. The analysis was carried out via the R package seqHMM (Helske and Helske, 2019).

As far as the second step is concerned, we employed a logistic regression to investigate the effect of users' prior behaviour on their likelihood of making a purchase. The logistic regression model (Cox (1958)) is defined as in Equation 2.

$$ln \frac{P(z_i = 1|W_i)}{P(z_i = 0|W_i)} = \beta W_i \quad (2)$$

Here, z_i represents the buying status (1 for purchase, 0 for no-purchase), β denotes the coefficient vector, and Wi consists of variables. The model is selected based on the BIC criterion and evaluated using sensitivity, specificity, and ROC-AUC measures.

3. Results and discussion

The empirical analysis has been performed on clickstream data collected from Sunweb.nl

website over the period May 24 to August 10, 2022.

At first stage, we have used MHMM to identify distinct behavioural clusters based on users' clickstream data. The model incorporated the access device (PC, mobile, or tablet) as a covariate variable and the best model was evaluated using the Bayesian information criterion (BIC), which determined that a model with K = 6 clusters was optimal. These six clusters, reflecting different browsing profiles, have been identified by different hidden Markov models. Furthermore, clusters' behaviour has been explored by analysing the subpatterns within each cluster, i.e., the short, consecutive, and frequently repeated subsequences of Web pages inside the sequences of the same cluster. We named the six profiles as: *Surfers & Employees* (Cluster 1) and *Surfers & Competitors* (Cluster 5) that exhibit repetitive browsing without clear purchasing intent; *Uncertain Explorers* (Cluster 4) focus on search and accommodation pages, while *Thoughtful Explorers* (Cluster 3) prioritise account and FAQ activity; *Travel Planners* (Cluster 6) show deliberate browsing of preferences and accommodation; *Confident Buyers* (Cluster 2) account for over 93% of purchases. Notably, Clusters 2 and 4 predominantly use desktop devices (80% and 67%), while others prefer mobile.

According to browsing patterns, these profiles indicate different stages in the customer journey, as they include users who are early-stage explorers, mid-stage planners, or the final-stage confident buyers mostly classified in cluster.

In the second stage, we focused on Cluster 2 which includes most purchasers, and we analysed the subset of traceable users that ended their Web journey with a visit classified in this cluster. These users spend the last visit either booking a holiday or leaving the website with no purchase event. To investigate the impact of previous behaviour on purchase likelihood, we applied logistic regression (Equation 2) using browsing behaviour profiles as covariates. Specifically, covariate matrix W_i includes explanatory variables such as total time spent on the website (time.spent), number of visits (num.vis), time between first and last visit (diff.time), and counts of visits classified into different behaviour profiles (count_BBP_c, for $c = (1, 2, \ldots, 6)$). The logistic regression model was optimised by minimising the BIC and was evaluated using oversampling techniques to balance the dataset's purchase and non-purchase statuses. The model achieved sensitivity of 0.66, specificity of 0.79, and an ROC-AUC of 0.80.

The analysis of marginal effects indicates that an increase in time.spent by one unit raises the probability of purchase by approximately 0.005%, while an additional visit (num.vis) decreases the purchase probability by about 0.685%.

Visits to Clusters 3 and 6 increase the purchase probability by 24.086% and 4.205%, respectively, and visits to Clusters 1 and 5 decrease the probability of purchase by 5.561% and 6.188%, respectively. Interaction terms show that users shifting between exploratory behaviour (Clusters 1, 3, and 5) and buying behaviour (Cluster 2) are less likely to make a purchase, whereas those moving from Explorers to *Confident Buyers* (Clusters 1, 3, 5 to 2) have a slightly increased purchase probability. These findings underscore that users are more inclined to purchase when they engage with the website in a focused manner rather than shifting behaviours frequently.

4. Concluding remarks

In conclusion, the study shows how traceable and untraceable information coming from clickstream data can be effectively used to model and predict online consumer behaviour in the tourism sector. By applying a mixture hidden Markov model to identify behavioural profiles and a logistic regression model to assess the impact of these profiles on purchase decisions, we gained valuable insights into user behaviour on the Sunweb.nl platform that can be used to tailor marketing strategies and to website design in order to engage users more effectively at various stages of their journey (Coussement et al. (2014)). Our findings indicate that users' varying browsing behaviours

during previous visits significantly influence their purchasing decisions. For example, users with a history of visits to certain clusters or longer engagement times show higher purchase intent.

In conclusion, our analysis highlights the potential benefits of implementing a real-time classification system for traceable users based on their behaviour after each site visit. By continuously assessing user profiles and their associated browsing patterns, the company could proactively engage with users, delivering personalized offers or targeted messages to increase their likelihood of making a purchase.

Finally, the study underscores the potential of advanced modelling techniques to provide actionable insights into online consumer behaviour, offering practical implications for businesses seeking to refine their digital strategies in highly competitive sectors.

References

Chen, S.X., Wang, X.K., Zhang, H.Y., Wang, J.Q., Peng, J.J. (2021). Customer purchase forecasting for online tourism: A data-driven method with multiplex behavior data. *Tourism Management*, **87**, 104357.

Coussement, K., Van den Bossche, F.A., De Bock, K.W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, **67**, pp. 2751–2758.

Cox, D.R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, **20**, pp. 215–232.

Dunn, M., Davis, S. (2003). Building brands from the inside. *Marketing Management*, **12**, 32–32.

Helske, S., Helske, J. (2019). Mixture hidden Markov models for sequence data: The seqHMM package in R. *Journal of Statistical Software*, **88**, pp. 1–32.

Huynh, H.M., Nguyen, L.T., Vo, B., Oplatková, Z.K., Fournier-Viger, P., Yun, U. (2022). An efficient parallel algorithm for mining weighted clickstream patterns. *Information Sciences*, **582**, pp. 349–368.

Kim, S., Shin, W., Kim, H.W. (2024). Predicting online customer purchase: The integration of customer characteristics and browsing patterns. *Decision Support Systems*, **177**, 114105.

Koehn, D., Lessmann, S., Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, **150**, 113342.

Moeller, S. (2010). Characteristics of services – A new approach uncovers their value. *Journal of Services Marketing*, **24**, pp. 359–368. Park, C.H.,

Park, Y.H. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, **35**, pp. 894–914.

Van de Pol, F., Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, pp. 213–247.

Vermunt, J.K., Tran, B., Magidson, J. (2008). Latent class models in longitudinal research. In: *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, pp. 373–385.

Ypma, A., Heskes, T. (2002). Automatic categorization of web pages and user clustering with mixtures of hidden Markov models. In: *International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles*, Springer, pp. 35–49.

Zhang, Y., Bradlow, E.T., Small, D.S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science*, **34**, pp. 195–208.

Zhou, Z. (2004). *E-commerce & Information Technology in Hospitality & Tourism*. Canada: Delmar Learning.

Zhu, G., Wu, Z., Wang, Y., Cao, S., Cao, J. (2019). Online purchase decisions for tourism e-commerce. *Electronic Commerce Research and Applications*, **38**, 100887.

doi.org/10.26398/asaproc.00142

Geography of business alliances and spatial network complexity: The backbone of formal collaboration agreements

Andrea Enrico Vurro[a], Alessio Bumbea[a], Annamaria Giuffrida[a], Andrea Mazzitelli[b], Giuseppe Espa[c]

[a] Department of Engineering and Sciences, Universitas Mercatorum, Rome, Italy.

[b] Department of Economics, Statistics, and Business, Universitas Mercatorum, Rome, Italy.

[c] Department of Economics and Management, University of Trento, Italy.

1. Introduction

To enhance the competitiveness of small businesses and foster enduring network relationships, the Italian government introduced in 2009 the *network contract*, a legal tool that governs and ensures strategic collaboration among firms (Gulati et al., 2000; Dickson et al., 2021).

A formal collaboration agreement (FCA) is a strategic choice for SMEs to form alliances with other companies. Early empirical studies have indicated that firms' involvement in network contracts is positively associated with the enhanced economic performance of firms (Burlina, 2020; Cisi et al., 2020). Notably, from analyses of the initial network contracts, smaller firms reaped the benefits of this instrument in terms of sales growth (Costa et al., 2017). Collaborative networks can help firms overcome strategic challenges and gain competitive advantages (Gulati et al., 2000).

These alliances allow firms to quickly access new assets, knowledge, capabilities, and technologies while maintaining flexibility and autonomy (Cohen and Levinthal, 1990; Latham and Le Bas, 2006). Increased access to external information and expertise (Malecki and Veldhoen, 1993) can strengthen a network member's market position or help enter new market segments (Laurell et al., 2017). However, measuring the long-term impacts of network involvement is difficult due to the scarcity of data. (Rosenfeld,1996). Since the legislation is relatively new, strategic business alliances generally require time to yield substantial benefits (Huggins, 2001), such as the absorption of new knowledge and improved learning capabilities. FCAs are crucial for SMEs as they enable firms to exchange information and experiences for mutual benefit by co-producing, co-marketing, co-purchasing, or collaborating on product or market development. (Dickson et al., 2021).

2. Data and variables

Our analysis, conducted with meticulous attention to detail, was based on a comprehensive dataset from the business register provided by InfoCamere, the Italian Chambers of Commerce IT company. Observing a direct relationship between networking and firm performance is difficult because performance is influenced by other factors that over time can produce value for the alliance between firms. These variables include the characteristics of the partners participating in a network agreement, such as the size of the network, geographic distance between firms (Cisi et al., 2020), i.e., geographic location, and sectoral diversity of its members (Schoonjans et al., 2013).

We specifically focused on network contracts without legal subjectivity, a type that is more prevalent than those with legal subjectivity. From this, we derived an initial sample of 5,607 network contracts without legal subjectivity that were active as of June 2024. Subsequently, we delved into the details of 3,035 network contracts, all of which involved firms with headquarters

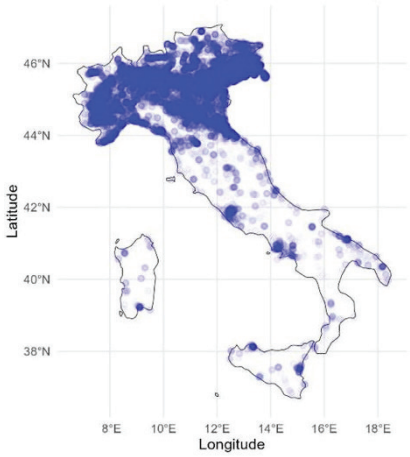
and local units in Northern Italy.

For each contract, the following information is available: the name of the contract, the date of joining the contract, the aims of the agreement, the tax codes, and the names of the participating enterprises.

The analysis is not just a surface-level examination. We have gone to great lengths to delve into additional demographic and economic information for the participating enterprises. This includes their regional (NUTS 2) and provincial (NUTS 3) location, prevailing economic activity (ATECO code), legal nature, date of birth of the enterprise, number of employees, and years of participation in the network contract. This comprehensive approach ensures a thorough understanding of the participating enterprises, providing a solid foundation for our research.

Thus, the sample includes about 11,300 of the 26,800 unterminated enterprises that have participated in at least one active network contract as of June 2024. For this sample, we considered not only the registered offices of the firms but also the local units, locating about 27,500 local units throughout northern Italy. We associated these local units with their corresponding four-digit ATECO codes (prevalent, primary, and secondary). By doing so, we constructed a dataset of approximately 54,700 local units with their corresponding geographic coordinates (Figure 1).

Figure 1 – Locations of the companies participating in FCA



Source: authors'elaboration

3. Methodology

In this section, we will introduce some measures of centrality that offer quantitative insight into the significance of a node within a network. Networks allow for the visualization of complex, multidimensional economic data and offer various statistical indices for interpreting the resulting graphs (Kolaczyk and Csárdi, 2020).

The structure we use to depict this is called a Bipartite graph, in which edges connect two entities. Only some connections in an extensive, complicated network are always equally significant. The most important connections are called the network's "backbone."

This 'backbone' structure is essential for figuring out meaningful connections and comprehending the network's fundamental structure.

The data on firms' geographic locations and ATECO code were converted into an undirected bipartite graph with 729 vertices and 9,827 edges. For our study, the main statistics computed on the bipartite graph examine centrality, and with the help of these, the backbone of the graph is computed. By removing some vertices, the structure of the graph G is simplified, resulting in a subgraph H that preserves the topology and most significant connections of the original graph, allowing for a simpler study.

In the context of graph theory and network analysis, the simplest measures to describe data

are the degree, the betweenness and the eigenvector centralities.

Degree Centrality, namely, the number of edges incident on a vertex, represents the sum of the number of edges passing by a vertex. Since the graph is undirected, then the degree of a vertex *v* is simply the number of edges of value 1 that connect it to the other vertices *u* within the adjacency matrix A according to the formula:

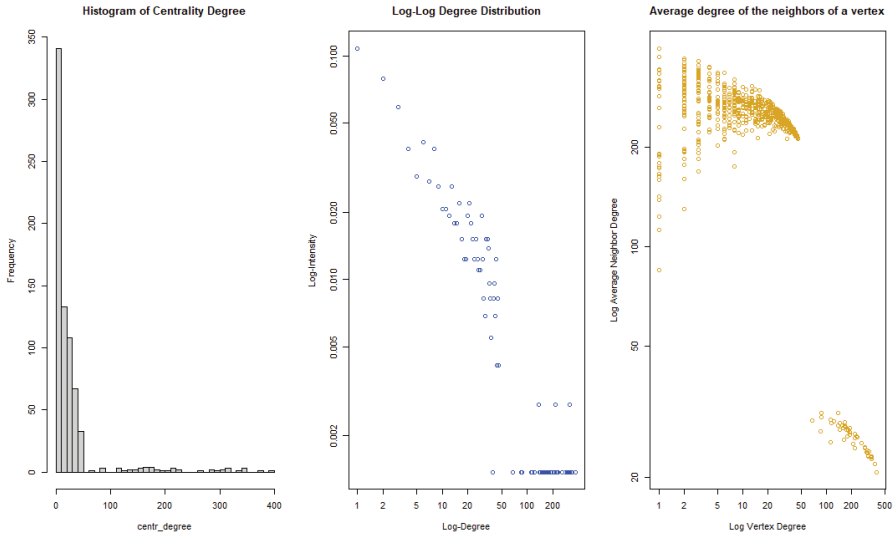
$$Degree(v) = \sum_{u \in V} A_{uv}$$

The higher the degree centrality of a node, the more connections the node has within the network. Nodes with a higher degree of centrality are more likely to be enabled for information and capabilities propagation.

The degree distribution of the vertices is shown below (Figure 2). The first graph shows the frequencies in a histogram, in the second the frequencies can be better appreciated thanks to a logarithmic scale, and in the third it is possible to observe the degree of neighboring vertices as a function of the degree of the vertex considered, i.e., vertices with low degree are connected to vertices with high degree and vice versa.

Higher-ranked nodes include the territories of Milano (398), Torino (374), Brescia (344), Udine (343) and Cuneo (342), while lower-ranked nodes include Lodi (68), Verbano Cusio Ossola (85) and Vicenza (86).

Figure 2 – The degree distribution



Source: authors' elaboration

Betweenness centrality was also calculated, which measures how far a node *v* lies on the minimum paths between other pairs of vertex's *s* and *t*, an indication of how much a node acts as a "bridge" that is, how essential it is to the structure of the graph. Nodes with high Betweenness are critical vertices; if these are missing, the graph comes could split into multiple components by not transmitting information.

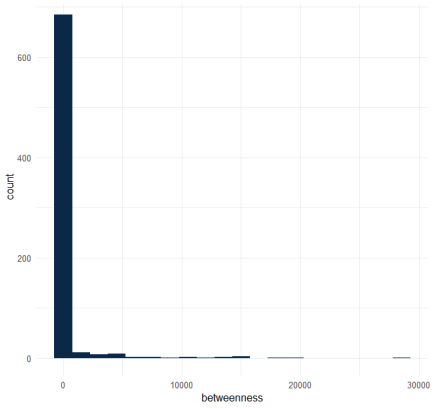
$$Betweenness(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st(v)}}{\sigma_{st}}$$

In the graph (Figure 3), the absolute frequency for values of betweenness shows the vertices with higher values: Milano (28533), Torino (20202), Cuneo (17383), Bergamo (15700) and Venezia (14985). Finally, eigenvector centrality was calculated, which measures not only the number of edges touching each vertex, but also the importance of the vertices to which that

vertex is connected; a node is considered important only if it is connected to other important nodes. The eigenvector centrality of a node *v* is an index between 0 and 1 defined as the *v-th* component of the principal vector of the equation $Ax = \lambda x$ where A is the adjacency matrix of the network (with $A_{ij} = 1$ if there is an arc between vertices i and j , otherwise with $A_{ij} = 0$), x is the vector of eigenvector centralities, where the *v-th* component x_v represents the eigenvector centrality of node *v* and λ is the largest eigenvalue of the matrix $\boldsymbol{A}$.

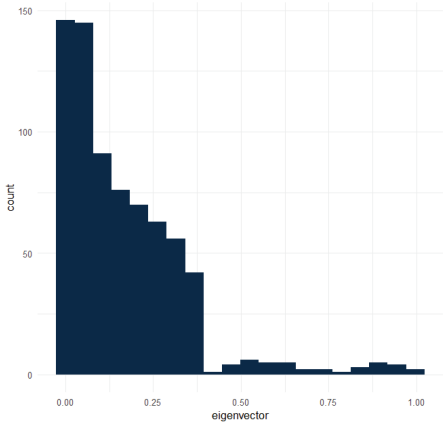
The greater the eigenvector centrality value of a node, the greater its importance within the network. The vertices with higher values represent the provinces of Milano (1.00), Torino (0.99), Udine (0.95), Brescia (0.95) and Cuneo (0.93) (Figure 4).

Figure 3 – Betweenness centrality distribution



Source: authors'elaboration

Figure 4 – Eigenvector centrality distribution



Source: authors'elaboration

4. The backbone model

Networks are valuable tools for modeling phenomena across various fields. While their ability to capture complexity is advantageous, simplifying a network to focus on its backbone can be beneficial in specific scenarios, such as computationally intensive analyses, dealing with noise, or improving visualization. For a complex network N, whether weighted or unweighted, its backbone N′ is a sparse, unweighted subgraph that aims to retain or highlight significant structural elements. Backbone models designed explicitly for bipartite projections are distinct because they operate directly on the bipartite network rather than on its projection. This approach allows these models to incorporate information inherent to the bipartite structure (Neal, 2021).

In a bipartite network, the vertices are separated into two partitions; each edge of the network connects the vertices of one partition to the vertices of the second partition, while there are no edges within the same partition. Therefore, the vertices V_1 of the first partition therefore are connected exclusively to the vertices V_2 of the second partition, and it is our interest to study the vertices of one partition pointing to the same vertices of the second partition, namely, the co-occurrences. Given the bipartite network $G_b = G(V_1, V_2, E_{1;2})$, we can proceed by multiplying the adjacency matrix B by its transposed matrix $P = BB^T$, thus projecting the bipartite network into a unipartite network $G_u = G(V_1, E_{1;1})$. In the resulting graph, vertices V_1 are found to be interconnected if they were connected to the same vertex V_2 in the bipartite network. To calculate the significant edges in *P*, we used the fixed degree sequence model (FDSM) for extracting the backbone from bipartite projections, which we implemented using the *fdsm*() function of Neal's (2021) backbone library. This function requires estimating *p-*

values using Monte Carlo simulation and it employs efficient methods for simulating random networks under the FDSM null model. Nevertheless, many simulations are still necessary to estimate *p*-values with sufficient confidence.

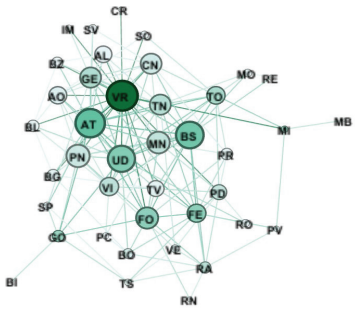
The FDSM is more appropriate than the hypergeometric model (HM) for most spatial data because it accounts for variations in rows and columns (Neal et al., 2022). FDSM is computationally intensive, but given the limited dimension of our dataset, it has been possible to use it proficiently. According to Neal (2021), other models exist, like SDSM, which is a close approximation of the FDSM.

Therefore, the factors guiding a choice between SDSM and FDSM backbones are not methodological, but practical (how large are the data?) and theoretical (how strict should the controls be?). When the data are small and/or strict control is desired, FDSM is more appropriate, while when the data are large and/or less strict control is suitable, SDSM is more appropriate**.** The resulting graph of the spatial backbone has been depicted in the following figures for different p-values (Figures 5-6). Particularly, vertices have size proportional to the eigenvector and colour intensity proportional to betweenness.

The fixed degree sequence odel application returned a p-value of 0.05 and a p-value of 0.10 for the networks with more components. These results are crucial as they guide our decision to retain vertices with degrees greater than zero, thereby shaping further analysis.

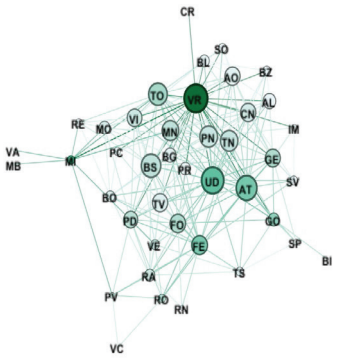
The graph with a p-value of 0.05 shows fewer focal nodes, suggesting a more selective network by retaining only the most significant connections; on the contrary, the graph with a p-value of 0.10 has more relevant nodes, indicating greater inclusiveness of connections, even the least significant ones; as a result, the eigenvector and betweenness values also vary slightly. The difference in *p*-values allows us to observe how the significance level affects density (0.146 applying a p-value = 0.05 versus 0.198 if p-value = 0.10) and network structure (average eigenvector of 0.338 using a p-value = 0.05 versus 0.387 if p-value = 0.10)

Figure 5 – Spatial network using backbone model – *p*-value =0.05



Source: authors'elaboration

Figure 6 – Spatial network using backbone model – *p*-value =0.10



Source: authors'elaboration

The analysis, conducted using the FDSM model, makes it possible to infer the spatial network among a set of locations (provinces) from data on their shared characteristics, i.e., the technological content of economic activities. This confirms the added value in terms of strategic alliance of geographic proximity among firms that have joined at least one network contract.

5. Conclusions

The present study explored the connections between ATECO codes and northern Italian provinces using the backbone of a graph with the FDSM model. The bipartite graph projection analysis revealed that some neighboring provinces show significant connections, suggesting a relevant correlation between geographic proximity and the distribution of economic activities

represented by ATECO codes. These results offer new insights into district and regional economic structure and the dynamics of interaction and centrality among provinces. Future studies could further explore these connections, examining the impact of specific ATECO subcategories and their evolution over time to understand better the economic and industrial dynamics of Italian provinces and businesses. The goal is to extend the study to the entire Italian territory to obtain a more complete view of economic connections at the national level.

In this work, we decided to limit our analysis scope to Northern Italy to reduce as much as possible issues related to spatial inhomogeneity.

References

Burlina, C. (2020). Networking policy and firm performance. *Growth and Change*, **51**(1), pp. 161–17.

Cisi, M., Devicienti, F., Manello, A., Vannoni, D. (2020). The advantages of formalizing networks: New evidence from Italian SMEs. *Small Business Economics*, **54**(4), pp. 1183–1200.

Cohen,W. M., Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, **35**(1), pp. 128–152.

Dickson, M. M., Espa, G., Gabriele, R., Mazzitelli, A. (2021). Small businesses and the effects on the growth of formal collaboration agreements: Additional insights and policy implications. *Applied Economics,* **53**(46), pp. 5397–5414.

Gulati, R., Nohria, N., Zaheer, A. (2000). Strategic networks. *Strategic Management Journal*, **21**(3), pp. 203–215.

Huggins, R. (2001). Inter-firm network policies and firm performance: Evaluating the impact of initiatives in the United Kingdom. *Research Policy*, **30**(3), pp. 443–458.

Kolaczyk, E.D., Csárdi, G. (2020). *Statistical Analysis of Network Data with R*, II edition, Springer

Latham,W.R., Le Bas, C. (2006). *The Economics of Persistent Innovation: An Evolutionary View*. Springer.

Laurell, H., Achtenhagen, L., Andersson, S. (2017). The changing role of network ties and critical capabilities in an international new venture's early development. *International Entrepreneurship and Management Journal,* 13(1), pp. 113–140.

Malecki, E. J., Veldhoen, M. E. (1993). Network activities, information and competitiveness in small firms. *Geografiska Annaler*: *Series B, Human Geography*, **75**(3), pp. 131–147.

Neal, Z.P. (2021) backbone: An R package to extract network backbones. *PLoS ONE* **17**(5), pp. 1-24.

Neal, Z.P., Domagalski, R., Sagan, B. (2022) Analysis of spatial networks from bipartite projections using R backbone package. *Geographical Analysis*, **54**(3), pp. 623-647.

Rosenfeld, S. A. (1996). Does cooperation enhance competitiveness? Assessing the impacts of inter-firm collaboration. *Research Policy*, **25**(2), pp. 247–263.

Rubino, M., Vitolla, F. (2016). The effects of network agreements on firms' performance. *Journal of Corporate Governance, Insurance and Risk Management,* **3**(1), pp. 17– 26.

Schoonjans, B., Van Cauwenberge, P., Bauwhede, H. (2013). Formal business networking and SME growth. *Small Business Economics*, **41**, pp. 169–181.

doi.org/10.26398/asaproc.00143

Spatial analysis of climate change awareness across countries

Gianpaolo Zammarchi[a], Paolo Maranzano[b]
[a] Department of Economics and Business Science, University of Cagliari, Cagliari, Italy.
[b] Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Milan, Italy.

1. Introduction

One of the greatest challenges that current and future generations will face is climate change and the extreme or severe repercussions this will entail. For instance, the increasingly frequent extreme weather events, the melting polar ice caps and warming oceans (with all the repercussions that these can have on the life of the animal species that live, hunt or base their life cycle on these places), and the rising sea levels are all effects of climate change. All countries are impacted by climate change, however certain areas may be more susceptible to particular effects. According to the Germanwatch observatory's annual assessment of the global climate risk index 2021, developing nations are more susceptible to the effects of climate change due to their reduced ability to adapt (Global Climate Risk Index, 2021). In fact, low- to lower-middle income countries account for eight of the ten most severely affected countries in terms of fatalities and economic losses caused by extreme weather events (such as storms, floods, heatwaves, etc.) in 2019. Three of the most affected countries were Mozambique, Zimbabwe and the Bahamas.

Increasing public awareness of the causes and effects of climate change is crucial because it can motivate policymakers to take action to cut greenhouse gas emissions and encourage individual behavioural adjustments. Since there are variations in how different nations might be vulnerable to some of the effects of climate change, the goal of this work was to examine how different nations differ in terms of their degree of knowledge of the problem.

Furthermore, we investigated the connection between awareness and pertinent economic indicators like GDP, Gini index, spending on environmental preservation, as these variables represent relevant characteristics of countries and present the preliminary results of an analysis based on a hierarchical clustering using spatial constraints.

2. Methodology

Two editions of the "International public opinion on climate change" survey were used to gather awareness data. This survey's main goals were to examine people's knowledge, beliefs, attitudes, policy preferences, and behaviour related to climate change. We collected data pertaining to the "climate awareness" question. In addition, we gathered data on GDP (GDP by Country, 2022) and environmental protection expenditures (Environmental Protection Expenditures, 2023), for the nations that were part of the study. Moreover, we created an extended version of the dataset with data collected from the Climate Change Knowledge Portal, the Penn World Table, the Global Carbon Atlas, the Standardized World Income Inequality Database, and the Our World in Data web portal, in order to be able to apply the spatially clustered methodology. Hierarchical clustering is a cluster analysis method that aims to create a hierarchy between clusters. There are two alternative approaches, that is, "bottom-up" (agglomerative) or "top-down" (divisive) and in both approaches

the concept of distance and its choice are of fundamental importance. The results of hierarchical clustering are usually shown in a graph called a dendrogram. Partitioning clustering, like K-means for instance, aims at dividing the data set into a specific number of clusters to minimize within-cluster variance while maximizing between-cluster variance. The algorithm assigns iteratively each of the *n* observations to one of the *k* clusters based on the distance between the observation and the cluster center. The main issue with K-means is that the number *k* of clusters is a hyperparameter that must be provided by the user.

In the first part of the analysis, we used data from the 2022 edition of the survey, including responses from 108,946 Facebook users located in 110 areas worldwide (Leiserowitz et al., 2022) as well as the 2021 edition of the survey, including responses from 76,328 users located in 30 countries (Leiserowitz et al., 2021). To investigate variations in climate change awareness across time, we incorporated information from the 30 nations that were included in both survey editions. Using the four response items, a pairwise Euclidean distance matrix was calculated in order to group nations based on their levels of awareness. As shown in Equation 1, single observations were combined into clusters using the Ward's approach.

$$\frac{|A| \cdot |B|}{|A \cup B|} \left|\left|\mu_A - \mu_B\right|\right|^2 = \sum_{x \in A \cup B} \left|\left|x - \mu_{A \cup B}\right|\right|^2 - \sum_{x \in A} \left|\left|x - \mu_A\right|\right|^2 - \sum_{x \in B} \left|\left|x - \mu_B\right|\right|^2 \qquad (1)$$

where A and B are two sets of observations.

Based on the elbow method, an optimal number of clusters was identified for both data sets. We created a hierarchical cluster using the hclust function in the *stats* R package (R Core Team, v. 4.3). The *dendextend* R package (Galili, 2015) was used to compare the two hierarchical clusters in order to highlight cluster variations that have occurred over the course of the two years. By matching the labels on the left side of the "double" hierarchical plot to the labels on the right, we were also able to calculate the "entanglement" value. The procedure is straightforward. Let v_1 and v_2 be two vectors corresponding to the order from 1 to *n* in which the labels appear on the left and right side of the plot, respectively. Moreover, let $\| \cdot \|_L$ be the L-norm distance between these two vectors. In order to obtain a maximum value *M* for this distance, the distance for the worst-case scenario is computed (which is when one side is the exact opposite of the other side). Finally, the "entanglement" can be written as: $E = \frac{||v_1 - v_2||_L}{M}$, with $E \in [0, 1]$. The two extreme scenarios, where $E = 0$ and $E = 1$, denote a perfect match between the two sets of labels and the scenario where the right tree is the left tree's full opposite, respectively. In addition, we applied K-means clustering to data from the 2022 edition of the survey with the goal of identifying groups of nations based on awareness, environmental protection expenditures, GDP, and knowledge of climate change. For this analysis, after merging and aligning data from different sources, we were able to use data from 62 countries.

The second part of the analysis is aimed at performing a hierarchical clustering with geographical constraints using an extended version of the dataset. In particular, we collected data from the above-mentioned sources to use the spatial clustering methodology proposed by Chavent et al. (2018). This method exploits two dissimilarity matrices called D_0 and D_1 along with a mixing parameter α that regulates the importance of the two matrices. The D_0 matrix provides information about dissimilarity in the "feature space", while the D_0 matrix in the "constraint space". Let I_α be the inertia, then the mixed pseudo inertia of cluster C_k^α can be defined as in Equation 2

$$I_\alpha(C_k^\alpha) = (1 - \alpha) \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i w_j}{2\mu_k^\alpha} d_{0,ij}^2 + \alpha \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i w_j}{2\mu_k^\alpha} d_{1,ij}^2 \qquad (2)$$

where w_i is the weight of the *i*-th observation for $i = 1,..., n$ which represents the importance of

observation i (similarly w_j denotes the importance of observation j), $\mu_k^\alpha = \sum_{i \in C_k^\alpha} w_i$ is the weight of C_k^α, $d_{0,ij}^2$ and $d_{1,ij}^2$ are the $n \times n$ normalized dissimilarity matrices between observations i and j in D_0 and D_1, respectively. Basically, the procedure performs a clustering based on the features collected to describe each country, while applying constraints (in this case spatial constraints) to create a combination that balanced the two aspects as a function of the value of α.

3. Results

In the first part of the analysis, we compared the two hierarchical clustering plots (2021 and 2022) to identify potential variations in the countries that were part of the clusters (Figure 1). The comparison of the two hierarchical clusters yielded an entanglement value of 0.29, indicating the existence of a discernible variation in the membership between the two years. Thailand and Colombia relocated to Cluster 3 from Cluster 2, i.e. in a cluster with poorer awareness of climate change, and Spain also experienced a decline in awareness, going from the cluster with the highest overall awareness to the intermediate one.

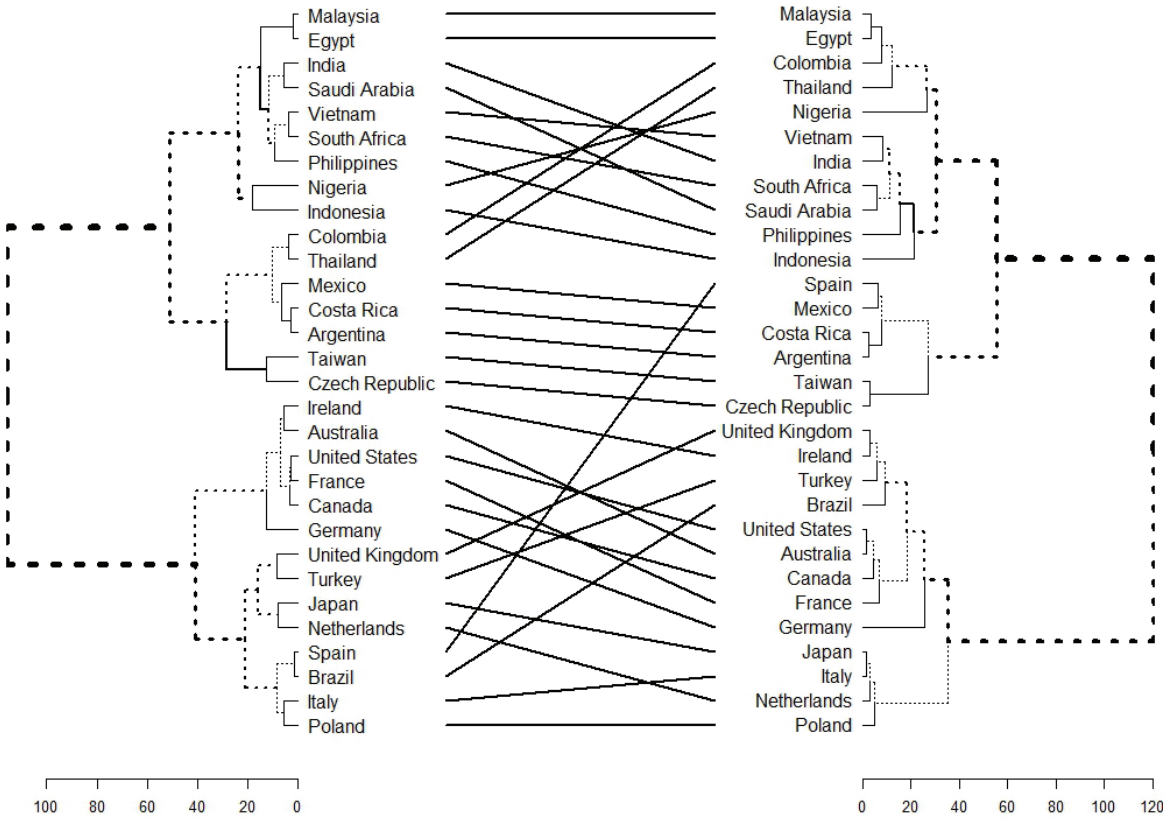


Figure 1. Comparison between awareness measured in 2021 and 2022

Using K-means, three clusters were identified based on the GDP, the environmental protection spending, and the levels of awareness: 1) countries with high awareness, GDP, and expenditure (green); 2) countries with low awareness, GDP, and expenditure (red); 3) countries with high GDP and awareness, but medium expenditure (Figure 2).

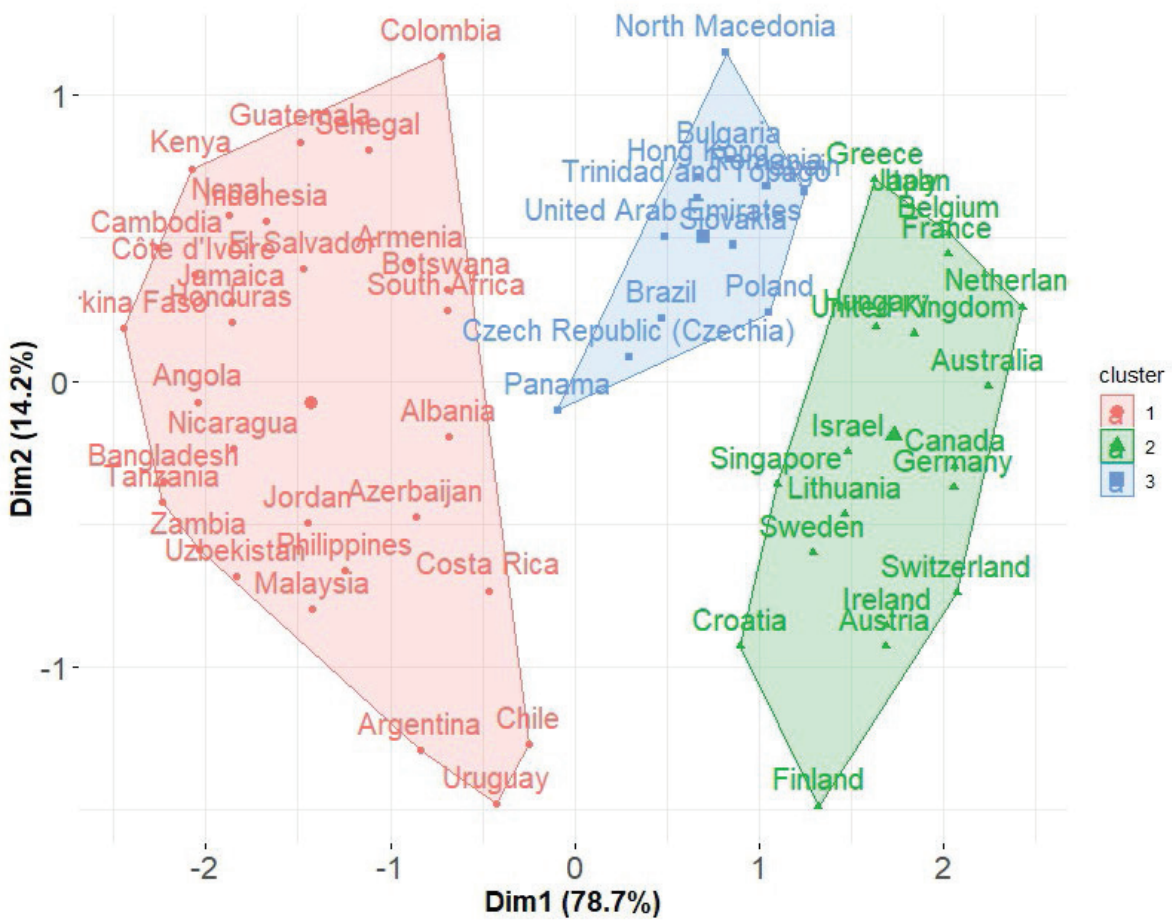


Figure 2. Clusters based on GDP, environmental protection spending, and awareness

Lastly, in Figure 3 we present a world map for all the countries analysed where the clustering was performed using the two matrices of features (based on awareness about climate change) and geographical dissimilarity. We selected an $\alpha = 0.55$ as a trade-off between D_0 and D_1 and set the number of clusters equal to 5. This figure shows that the geographical component is predominant, but nevertheless in some cases countries were not included in the most expected geographical cluster, but in another (e.g. Spain was not included in the European cluster but in the African one). Furthermore, Figure 4 provides additional information on the mean awareness levels of the countries in each cluster.

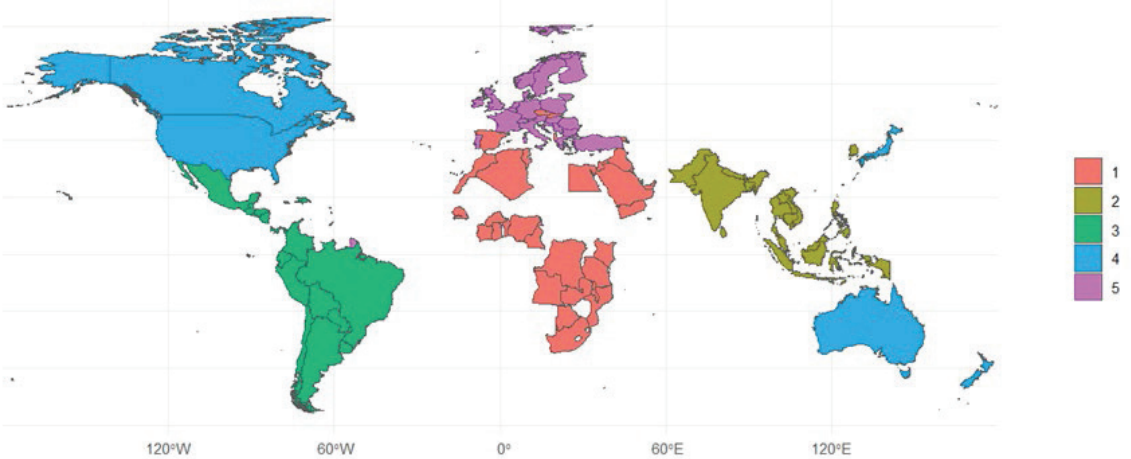


Figure 3. Clusters based on D_0 and D_1 matrices of features and geographical dissimilarities

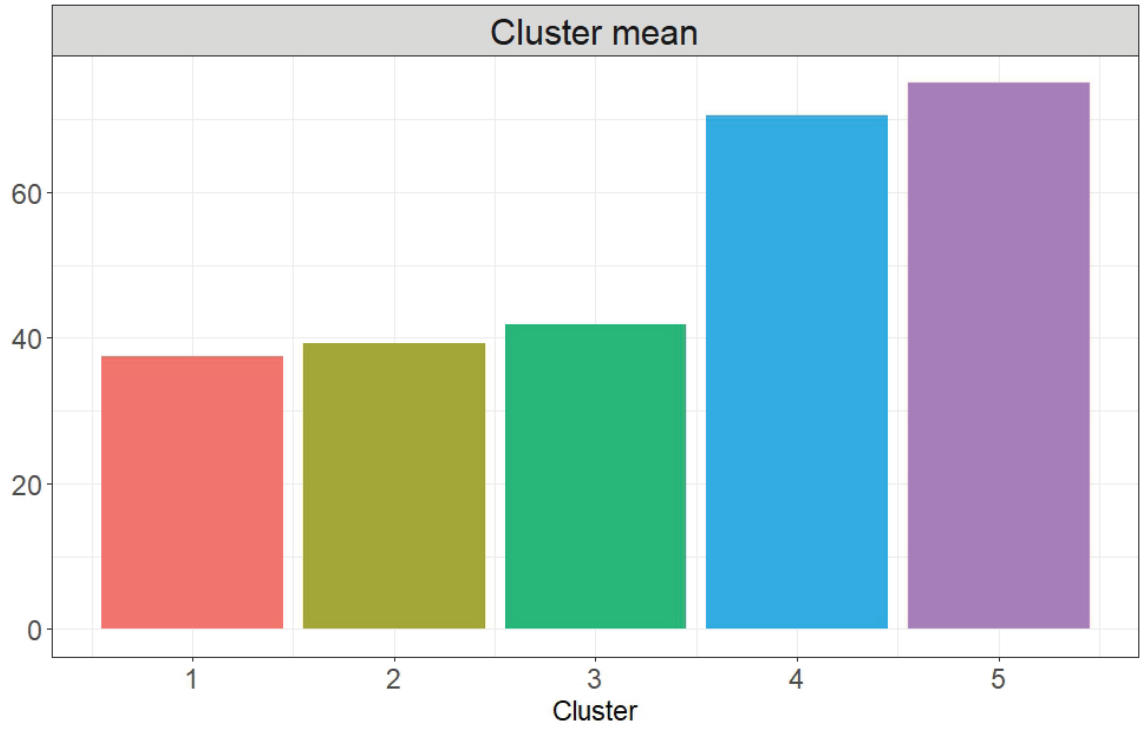


Figure 4. Barplot of awareness levels for the clusters created combining the D_0 and D_1 matrices

4. Conclusion

In this preliminary analysis, we studied the relationship between public awareness of climate change and variables such as GDP and environmental protection spending for each country considered. We performed a cluster analysis of nations according to their awareness of climate change and created three groups using economics and environmental variables. Finally, we produced an extended version of our dataset to be able to find the best value for α and then we used that value to create clusters based on both the features dissimilarity matrix and the geographical dissimilarity matrix. The clustering results consistently demonstrate that countries with high awareness levels are the richest countries (mostly Europeans and north Americans), while countries with low awareness are south Americans, Africans and part of the Asian. These countries could focus on educational programs and might share resources and plan strategies together to better address the problems related to increasing climate change awareness levels.

Future improvements of this work will include to expand the analysis and to explore the spatial correlations between countries in order to assess whether geographically neighbouring nations share a comparable level of awareness on climate change.

References

Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J. (2018). ClustGeo: An R package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4), pp. 1799-1822.

Climate Change Knowledge Portal (2024). Last access: July 31, 2024, https://climateknowledgeportal.worldbank.org/

Environmental Protection Expenditures (2023). Last access: December 15, 2023, https://climatedata.imf.org/datasets/d22a6decd9b147fd9040f793082b219b/explore

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), pp. 3718-3720.

GDP by Country (2022). Last access: December 15, 2023, https://www.worldometers.info/gdp/gdp-by-country/

Global Carbon Atlas (2024). Last access: July 31, 2024, https://globalcarbonatlas.org/emissions/carbon-emissions/

Global Climate Risk Index (2021). Last access: December 15, 2023, https://reliefweb.int/report/world/global-climate-risk-index-2021.

Leiserowitz, A., Carman, J., Buttermore, N., Wang, X., Rosenthal, S., Marlon, J., Mulcahy, K. (2021): *International Public Opinion on Climate Change*. New Haven, CT: Yale Program on Climate Change Communication and Facebook Data for Good .

Leiserowitz, A., Carman, J., Buttermore, N., Neyens, L., Rosenthal, S., Marlon, J., Schneider, J., Mulcahy, K. (2022): *International Public Opinion on Climate Change*, New Haven, CT: Yale Program on Climate Change Communication and Data for Good at Meta.

Our World in Data (2024). Last access: July 31, 2024, https://ourworldindata.org/

Penn World Table (2024). Last access: July 31, 2024, https://www.rug.nl/ggdc/productivity/pwt/?lang=en

R Core Team (2023). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. http://www.R-project.org

Standardized World Income Inequality Database (2024). Last access: July 31, 2024, https://fsolt.org/swiid/

doi.org/10.26398/asaproc.00144

Future scenarios: A financial network analysis of stock prices in emerging metaverse market

Emma Zavarrone [a], Alessia Forciniti [a]
[a] Department of Humanities, IULM University, Milan, Italy.

1. Introduction

The current digital landscape is undergoing a metamorphosis, driven by the convergence of social interaction and emerging technologies as the *metaverse*, defined as "an interconnected web of ubiquitous virtual worlds partly overlapping with and enhancing the physical world" (Weinberger, 2022). The metaverse, a persistent virtual space, is gaining significant attention. This concept has the potential to revolutionize how we socially interact and work. Its development is driven by technological advancements and industry interest, particularly from companies like Meta. As the metaverse evolves, it offers numerous investment opportunities across various sectors, including virtual economies, customer engagement, and operational efficiency.

This paper explores the financial implications of the metaverse on stock prices of companies involved in its development. It investigates the network-based correlation between the *Metaverse index* (MVI) and daily stock prices of metaverse-oriented companies. The MVI, a pool of 15 decentralized tokens, serves as a benchmark for tracking the growth of the metaverse. Two key questions are addressed in this study:

RQ_1: Is there a correlation between the MVI and the market capitalization of metaverse-oriented companies?

RQ_2: How does the relationship between the MVI and stock prices of metaverse-oriented companies evolve over time?

By exploring these questions, this work sheds light on the financial prospects of the metaverse and provides valuable insights for investors and stakeholders in this burgeoning field. Our research builds upon Bushnell's (2022) work on the interconnectedness of cryptocurrencies and other financial assets, aiming to identify patterns between the MVI and stock prices. By analyzing this relationship, we seek to understand whether the metaverse hype is primarily driven by speculation or reflects a genuine correlation with the value of other assets.

2. Methods

Financial markets are intricate systems with complex interconnections (Mantegna and Stanley, 2000). This study employs a network-based approach to analyze these relationships. By examining correlations between stock market data, we aim to uncover hidden patterns and structures within the financial network. Three key steps inspired by Katsiampa et al. (2022) represent the methodology: *i)* stock market network construction and filtering, *ii)* stock market network clustering, and *iii)* stock market prediction. In detail:

i) Stock price data underwent log return transformation and were used to construct a network-based correlation for each year examined. To capture the true structure of the network based only the strong links, *minimum spanning tree* (MST) and *triangulated maximally filtered graph* (TMFG) (Massara *et al.,* 2017) community detection algorithms were performed.

ii) To evaluate the best partitioning between MST and TMFG, the concept of *modularity* has been applied. As optimization of modularity, the Q metric was involved. It measures the difference between the fraction of edges within communities and the expected fraction of edges within communities in a random graph with the same degree distribution (Newman and Girvan, 2004). Q ranges between 0 and 1, with empirical optimal values between 0.3 and 0.7.

The core representation of our daily stock price network is the adjacency matrix (Wassermann and Faust, 1994) that records the connections between nodes or vertex (i.e., daily stock prices) with corresponding labels appearing on both rows and columns. Each entry reflects the strength of the connection between two nodes (Everett and Borgatti, 2020). In our analysis, the correlation matrix of daily stock price serves as the foundation for: the adjacency matrix, network construction and metrics computation. Numerous network metrics were used to analyze the network structure. To delve deeper into the complex interconnections, measures like those based on shortest distances were crucial.

iii) To explore potential future scenarios within these communities, a probabilistic predictive model was developed, after identifying the best communities. This model leverages node embedding techniques to represent nodes as vectors in a multi-dimensional space. Node embeddings are low-dimensional vector representations of nodes in a graph. These vectors capture the structural information of nodes within the network. By mapping nodes to dense vectors, node embeddings allow us to apply various machine learning techniques to graph data. In this study, the *node2vec* algorithm (Grover and Leskovec, 2016), which involves optimizing a neighborhood-preserving objective, was adopted.

The forecasting future trends in the metaverse, characterized by volatility and uncertainty, is challenging. Combining probabilistic forecasting with node2vec embeddings can enhance predictive accuracy. The node2vec can capture complex relationships, providing valuable input for probabilistic models. This integration can lead to more accurate predictions of future events and trends within the metaverse ecosystem.

3. Data acquisition and assets selection

A daily stock price time-series of MVI and 47 metaverse-oriented assets were retrieved from Coinbase marketplace for the period 2019 to 2023, comprising 1,827 observations for 48 variables. Data extraction satisfied Bloomberg criteria, specifically targeting:

- Companies deriving greater than 10% of their revenue from the metaverse business line;
- Companies with an estimated growth exceeding 10% in the metaverse sector by 2025;
- Companies allocating more than 10% of their capital expenditure to the metaverse;
- Companies acting as suppliers or solution providers for businesses active in the metaverse.

The retrieved assets were categorized in 13 areas as follows:

a) Processing & IT / Operating Systems: Broadcom, Nvidia, Alphabet, Meta, Texas Instruments (TI); b) Networks & Fixed Connection: Amazon, Foxconn, Fastly; c) Platforms & platform development: Nintendo, Microsoft, Unity, Apple; d) Interchange: Adobe; e) Payments: Mastercard, Visa; f) Content: Cosmos, Tencent, CapCom, Nike, NetEase; g) Cryptocurrencies: Bitcoin, Ether, Cardano, Solana, Dogecoin; h) Safe Havens: Gold, Crude Oil (COil), Natural Gas (NGas), Silver; i) Consumer staples: McDonalds, P&C, Unilever, Tesco, Coke; j) Real estate: Crown Castle (CCastle), US RE, Euro RE, Developed Markets Property (DMP); k) Treasuries: US Dollar (USD), German Bund (GB); l) Volatility indices: VIX and VNX; m) MVI ETFs: (CMC Crypto 200 Index by Solactive, Horizon Metaverse Index (HMI), Roundhill Ball Metaverse ETF, and ProShares Metaverse ETF).

4. Results and conclusions

With regards to the methodological approach *i*), the analysis of daily stock price time series from 2019 to 2023 using conventional linear correlation proved the lack of correlation between the MVI and commonly recognized assets within the digitalized market. Linear correlations among assets within the same domain were detected, aligning with existing literature (e.g., Momtaz, 2022). The weak or absent correlation between the MVI and other assets, including market volatility indices, can

potentially be attributed to the nascent stage of the metaverse.

Network structures analyzed using both the MST and TMFG methods (methodological approach ii)) revealed a transition in network partitioning from cohesive relationships, representing shared market segments, to a more hybridized state, suggesting an ongoing progression. In details, the long-term network analysis, spanning from 2021 to 2023, returned interesting connections of MVI. The findings outline an evolution of the metaverse towards new fields such as virtual economies, real estate markets, content, and entertainment. The results of community detection are visualized (Figure 1), using a Sankey diagram. This diagram provides a visual summary of the metaverse's market-driven companies over time.

Notably, the metaverse's evolution has led to increased interconnectivity among diverse sectors. The volatility-indices of the financial markets (i.e., VIX, VXN), and IT companies exhibit greater stability and less movement across sectors in the period examined. As the metaverse consolidated over time, stronger relationships (defined in under network 'perspective) emerged, fostering greater cross-sector synergy within the market. For example, cryptocurrencies, which were fragmented across various groups in 2019, demonstrate an openness towards the metaverse-assets in 2023. This is evidenced by virtual economies being clustered together and associated with content and metaverse indices such as CMC200. Similarly, the evolution of real estate alongside METV emphasizes the rise of virtual property. The MVI reflects this shift, prioritizing assets related to cryptos, volatility, content, and operating systems. The MVI's lack of centrality suggests a broader vision for the metaverse, encompassing not just digital economies but also entertainment, storytelling, and user experience.

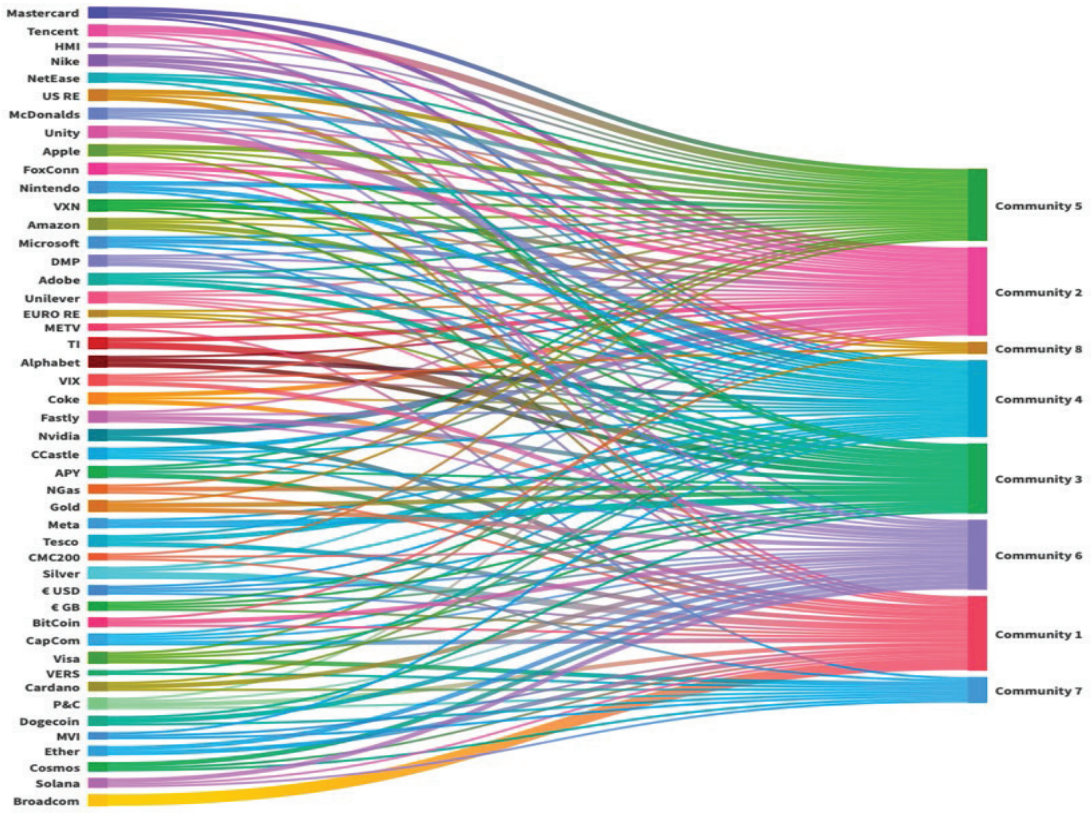


Figure 1: Sankey diagram of long-term scenario: from 2019 to 2023

The better performances, in terms of Q measure, are related to short-period analyses TMFG than MST, where for TMFG $Q > 0.70$.

Thus, for methodological approach *iii)* based on future scenario prediction, the node2vec

algorithm was applied on TMFG 2022. To determine the best node2vec model, various combinations of its parameters were performed on 2022 data to forecast 2023 data, while keeping the error between the predicted and observed links under control. The chosen model was the one that minimized the error and maximized the overlap between the real and predicted links.

The set of node2vec parameters identified to predict 2023 based on 2022 data was applied to 2023 data to forecast 2024. The result of embedding for determining future scenario 2024 is summarized into Table 1.

The forecasting 2024 on embedding of 2023 data shows technological advancements that are rapidly shaping the metaverse landscape in this last year.

This transformation impacts consumer behavior and social interactions, fostering the emergence of novel connection methods within virtual environments. Digital ownership, virtual economies, and cryptocurrencies, defining the meta-economy, are poised for further growth. Decentralized finance (DeFi) has gained prominence. The MVI, while not central, connects various sectors including entertainment, content, and payments. As the metaverse matures, the relationship between traditional finance and the metaverse may stabilize. Entertainment and DeFi are expected to be more developed in 2024, while VIX and VXN may no longer be directly linked to metaverse fluctuations in a more mature market. They could still serve as indicators for the emerging metaverse in the long run. The MVI is expected to exhibit a reduced correlation with traditional volatility (VIX) due to the metaverse's diversification and integration with sectors like real estate. Overall, the research highlights the evolving nature of the metaverse and its impact on traditional financial markets. By navigating the metaverse financially, short-term network analysis revealed the TMFG method as the best approach to capture the sector's transformation within the metaverse. This analysis identified a strong connection between the MVI, new technologies, and meta-economy.

Table 1: Forecasting 2024 on embedding 2023 data

Assets	Community	Assets	Community
Amazon, Broadcom, P&C, Tencent, TI, VIX	**1**	Fastly, Silver, Coke, Cosmos, NetEase, Unilever, METV, MVI, Bitcoin	**4**
Adobe, Apple, VXN, CMC200, HMI, DMP, Nintendo, Nvidia, Visa, Coil	**2**	McDonalds, Dogecoin, Meta, CapCom, NGas, VERS, Mastercard, Cardano	**5**
CCastle, Nike, Ether, Solana, Gold, US RE, Microsoft, Tesco, €GB, FoxConn	**3**	APY, Unity, EURO RE, Alphabet, €USD	**6**

Future scenarios highlight limitations related to hardware and software, hindering seamless interoperability between user platforms, a key aspect of a truly interconnected metaverse. The MVI also faces challenges due to low connectivity within sectors. To address these issues, developing metaverse-native companies offering innovative virtual goods and services is crucial. These companies should focus on intellectual property to become significant and innovative assets, potentially influencing the stock market landscape. New investment strategies and asset categories might emerge, requiring new stock market indexes.

Among the limitations of this work, which are taken into account in drawing conclusions, there is the brevity of the time series, which is, however, connected to the characteristic of the emerging

phenomenon of the metaverse. The short time series available for MVI limits to draw predictions about long-term trends, but provides valuable insights into short-term trends. For long-term forecasts, we rely on minimizing the error in node2vec model.

With reference to the methodological approach, while remaining within the framework of node embedding, other algorithms can be considered that overcome the limitations of node2vec known in the literature and may present different results and performance. As alternative methodologies, there are further approaches aimed at studying the proximity between nodes, expanding the dataset. Finally, node2vec could be applied to non-linear correlation matrices to explore possible different link dynamics between nodes.

References

Bushnell, M. (2022). Global capital markets and cryptocurrency: Exploring the international political economy of blockchain ecosystems and metaverse development. *Advance,* Sage Preprint. https://advance.sagepub.com/doi/full/10.31124/advance.19742272.v1

Everett, M.G., Borgatti, S.P. (2020). Partitioning multimode networks, in P. Doreian, V. Batagelj, and A. Ferligoj, eds. *Advances in Network Clustering and Blockmodeling*, Wiley, pp. 251-265.

Grover, A., Leskovec, J. (2016). node2vec: Scalable feature learning for networks, in *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco, California.

Katsiampa, P., Yarovaya, L., Zieba, D. (2022). High-frequency connectedness between bitcoin and other top-traded crypto assets during the COVID-19 crisis. *Journal of International Financial Markets, Institutions and Money*,**79**, 101578.

Mantegna, R.N., Stanley, H.E. (2000). *An Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge University Press.

Massara, L., Previde Massara, G., Di Matteo, T., Aste, T. (2017). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Network*, **5**, pp. 161-178.

Momtaz, P.P. (2022). Some very simple economics of Web3 and the metaverse. *FinTech.*

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

Wasserman, S., Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Weinberger, M. (2022). What is metaverse? - A definition based on qualitative meta-synthesis. *Future Internet*, **14**.