

Italian Statistical Society Series on Advances in Statistics

Alessio Pollice · Paolo Mariani *Editors*

Methodological and Applied Statistics and Demography IV

SIS 2024, Short Papers,
Contributed Sessions 2



**Italian Statistical Society Series on
Advances in Statistics**

This book series publishes volumes of peer-reviewed short papers presented at the scientific events such as conferences, seminars and workshops organized by the Italian Statistical Society (SIS) and its sections.

The Italian Statistical Society (Società Italiana di Statistica, SIS) was established in 1939 and ranks among the institutions of particular scientific relevance. SIS aims to promote scientific activities for the development of statistical sciences and carries out this task organizing scientific meetings and conferences, and by means of publications and partnerships at a national and international level.

Alessio Pollice · Paolo Mariani
Editors

Methodological and Applied Statistics and Demography IV

SIS 2024, Short Papers, Contributed Sessions 2

Editors

Alessio Pollice
Department of Economics and Finance
The University of Bari Aldo Moro
Bari, Italy

Paolo Mariani
Department of Economics, Management
and Statistics
University of Milano-Bicocca
Milan, Italy

ISSN 3059-2135

ISSN 3059-2143 (electronic)

Italian Statistical Society Series on Advances in Statistics

ISBN 978-3-031-64446-7

ISBN 978-3-031-64447-4 (eBook)

<https://doi.org/10.1007/978-3-031-64447-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The promotion of scientific activities for the advancement of statistical sciences represents the defining element of the Italian Statistical Society's (SIS) actions. Since its inception, the society has grown in tandem with the rapid dissemination of quantitative methods for data analysis across various fields of scientific research and social life. Academics, researchers, doctoral students, and professionals interested in statistical and demographic research, as well as related methodological and applied fields, have shared their activities during the 52nd Scientific Meeting of the Italian Statistical Society held at the University of Bari Aldo Moro from June 17 to June 20, 2024.

The biennial conference is a traditional event that promotes interactions among national and international researchers in statistics, demography, and applied statistics in Italy. The objective of the conference is to bring together national and foreign researchers and professionals to discuss recent developments in theoretical and applied statistics, as well as demography and statistics for the social sciences. The event featured over 450 presentations, including 7 keynotes in the 5 plenary sessions, 66 contributions in the 22 specialized sessions, 144 contributions in the 36 invited sessions, and 236 presentations in the 48 sessions of spontaneous communications, all addressing specific themes of methodological and/or applied Statistics and Demography. The high number of contributions and the significant participation in the conference demonstrate the desire to participate in scientific events as venues for exchange and discussion on the new developments in our field, a vibrant characteristic of our scientific community.

Round tables and meetings on topics of general interest such as “Data science: statistics for the 22nd century” and “Policy tools for innovation and development” were also organized, along with satellite events such as the 7th edition of the hackathon “Stats Under the Stars”.

We would also like to thank the European Network for Business and Industrial Statistics (ENBIS) and the Institute of Mathematical Statistics (IMS) for their presence. Special thanks goes to the program and organizing committees for their tremendous effort, to the University of Bari Aldo Moro for endorsing and supporting the event, and to all the sponsors for their contributions: Società Italiana di Statistica, Università degli Studi di Bari Aldo Moro, Banca Popolare di Puglia e Basilicata, Fondazione Puglia, Ladisa Ristorazione S.R.L., and Unioncamere Puglia.

The high number of contributions has led to the production of four volumes based on the types of sessions outlined in the scientific program. The first volume includes the plenary and specialized sessions, the second volume includes the solicited sessions, and the third and fourth volumes include the contributed sessions. These volumes contain works, as usual, subjected to peer review, that the authors have agreed to publish, providing a comprehensive overview of statistical research in recent years.

This fourth volume presents a selection of contributions from authors who delivered presentations during the contributed sessions.

June 2024

Alessio Pollice
Paolo Mariani
Editor and Chair of the Organizing
Committee, SIS 2024

Organization

SIS 2024 is organized by the department of Economics and Finance, University of Bari Aldo Moro in cooperation with the Italian Statistical Society.

Editors

Alessio Pollice	University of Bari Aldo Moro
Paolo Mariani	University of Milano, Bicocca

Executive Editors

Diego Battagliese	University of Bari Aldo Moro
Massimo Bilancia	University of Bari Aldo Moro
Crescenza Calculli	University of Bari Aldo Moro
Vittorio Nicolardi	University of Bari Aldo Moro

Organizing Committee

Alessio Pollice (Conference Chair)	University of Bari Aldo Moro
Diego Battagliese	University of Bari Aldo Moro
Antonella Bianchino	Italian National Institute of Statistics
Massimo Bilancia	University of Bari Aldo Moro
Crescenza Calculli	University of Bari Aldo Moro
Francesco Campobasso	University of Bari Aldo Moro
Maria Carella	University of Bari Aldo Moro
Carlo Cusatelli	University of Bari Aldo Moro
Francesco Domenico D'Ovidio	University of Bari Aldo Moro
Angela Maria D'Uggento	University of Bari Aldo Moro
Najada Firza	University of Bari Aldo Moro
Michele Gallo	University of Napoli, L'Orientale
Thais Garcia Pereiro	University of Bari Aldo Moro
Samuela L'Abbate	University of Bari Aldo Moro
Fabio Manca	University of Bari Aldo Moro
Claudia Marin	University of Bari Aldo Moro

Caterina Marini	University of Bari Aldo Moro
Antonella Massari	University of Bari Aldo Moro
Rocco Mazza	University of Bari Aldo Moro
Dante Mazzitelli	University of Bari Aldo Moro
Roberta Misuraca	University of Bari Aldo Moro
Vittorio Nicolardi	University of Bari Aldo Moro
Roberta Pace	University of Bari Aldo Moro
Anna Paterno	University of Bari Aldo Moro
Michela Camilla Pellicani	University of Bari Aldo Moro
Paola Perchinunno	University of Bari Aldo Moro
Nunziata Ribecco	University of Bari Aldo Moro
Ernesto Toma	University of Bari Aldo Moro
Giovanni Vannella	University of Bari Aldo Moro
Domenico Viola	University of Bari Aldo Moro

Program Committee

Paolo Mariani (Conference Chair)	University of Milano, Bicocca
Elena Ambrosetti	University of Roma, Sapienza
Laura Anderlucci	University of Bologna
Margaret Antonicelli	University of Roma, Sapienza
Raffaele Argiento	University of Bergamo
Maria Gabriella Campolo	University of Messina
Maurizio Carpita	University of Brescia
Enrico Ciavolino	University of Salento
Andrea Ciccarelli	University of Teramo
Enrico Di Bella	University of Genova
Giuseppe Gabrielli	University of Napoli, Federico II
Salvatore Ingrassia	University of Catania
Antonio Irpino	University of Campania
Fulvia Mecatti	University of Milano, Bicocca
Stefano Menghiniello	Italian National Institute of Statistics
Alessio Pollice	University of Bari Aldo Moro
Mariangela Sciandra	University of Palermo
Giorgia Zaccaria	University of Roma, Sapienza

Referees

G. Adelfio	G. Gabrielli
L. S. Alaimo	M. Gallo
A. Albano	G. Ghellini
G. Alderotti	A. Ghiglietti
M. Angelelli	F. A. Giambona
R. Argiento	C. Gigliarano
S. Arima	S. Giordano
F. Ascolani	C. Gozzi
D. Battagliese	A. Guglielmi
F. Benassi	M. Iannario
M. Beraha	R. Ignaccolo
R. Berni	R. Leombruni
C. Bernini	C. Liberati
G. Bertarelli	B. Liseo
I. Bombelli	R. Lombardo
C. Bonifazi	A. Lucadamo
A. Busetta	F. Maggino
G. Busetta	C. Marin
A. Calcagni	A. Marletta
C. Calculi	F. Martella
M. Cameletti	G. Mastrantonio
F. Camerlenghi	A. Mattei
M. G. Campolo	A. Mazza
A. Canale	S. Menghinello
C. Capezza	C. Mollica
A. Cappozzo	A. Naccarato
M. Carpita	M. Nai Ruscone
A. Casa	B. Nipoti
A. Cassese	M. Novelli
C. Castagnaro	R. Pace
C. Cavicchia	S. Pacei
R. Cerqueti	L. Paci
A. Ciccarelli	A. Paparusso
C. Conversano	F. Pavone
A. D'Ambra	C. Polli
A. D'Ambrosio	A. Pollice
C. Di Serio	M. Pratesi
S. Di Zio	I. Prosdocimi
F. D'Ovidio	G. Punzo
A. M. D'Uggento	G. Ragozini
E. Fabrizi	F. Rapallo
F. Finazzi	A. Rinaldi
S. Franceschi	E. Ripamonti

E. Romano	E. Trappolini
S. Salini	C. Valentino
L. Sangalli	R. Verde
P. Sarnacchiaro	P. Vicard
M. Sciandra	C. Viroli
A. Simonetto	M. P. Vitale
S. D. Tomarchio	M. Zenga
V. Tomaselli	

Sponsoring Institutions

Società Italiana di Statistica, Roma, Italy
Università degli Studi di Bari Aldo Moro, Bari, Italy
Banca Popolare di Puglia e Basilicata, Altamura (BA), Italy
Fondazione Puglia, Bari, Italy
Ladisa Ristorazione S.R.L., Bari, Italy
Unioncamere Puglia, Bari, Italy

Contents

Spatial Ranking Analysis on Quality of Services Using Some Techniques to Improved Fuzzy c-Means Algorithm	1
<i>Najada Firza, Laura Antonucci, and Dante Mazzitelli</i>	
A Topology-Based Algorithm for the Isomorphism Check of 2-Level Orthogonal Arrays	9
<i>Roberto Fontana and Marco Guerra</i>	
The Revolution of AI in Healthcare Between Diagnosis and Treatment	15
<i>Alessia Forciniti and Francesco Santelli</i>	
Hidden Markov Graphical Models with Generalized Hyperbolic Distributions: A Financial Analysis on Commodities and Green Energy Indexes	22
<i>Beatrice Foroni, Luca Merlo, and Lea Petrella</i>	
L_0 -Penalized Membership in Sparse Fuzzy Clustering	28
<i>Maria Brigida Ferraro, Marco Forti, and Paolo Giordani</i>	
Secondary Migration Intentions of Refugees Living in Germany	33
<i>Cecilia Fortunato, Hans Dietrich, and Elena Ambrosetti</i>	
A Bayesian Binomial Regression Model for Ozone Levels in Northern Italy ...	40
<i>Michela Frigeri, Leonardo Marchesin, Matteo Coppellotti, and Alessandra Guglielmi</i>	
Adaptive Elastic-Net Estimation for Ergodic Diffusion Processes	46
<i>Dario Frisardi, Alessandro De Gregorio, Francesco Iafrate, and Stefano M. Iacus</i>	
Internet Activities and Life Satisfaction Among Older People in Italy	52
<i>Claudia Furlan and Silvia Meggiolaro</i>	
Integrating Pseudo-time in Network Interaction Analysis to Uncover Genetic Pathways in Thyroid Cancer	58
<i>Antonino Gagliano, Gianluca Sottile, Nicolina Sciaraffa, Claudia Coronello, and Luigi Augugliaro</i>	
Housing Insecurity and First Births in Italy	64
<i>Alessandro Gallo and Daniele Vignoli</i>	

Investigating Auxiliary Mixture Sampling for Poisson Regression Models	71
<i>Aldo Gardini, Fedele Greco, and Carlo Trivisano</i>	
Unobserved Heterogeneity and Identification of Causal Effects Using Mixture Models	77
<i>Leonardo Genesin, Marco Bertoni, Adriano Paggiaro, and Giovanna Menardi</i>	
A Flexible Hybrid Method for Sample Size Calculations in Comparative Poisson Trials	83
<i>Susanna Gentile and Valeria Sambucini</i>	
Combining Bootstrap Methods: A Monte Carlo Experiment	90
<i>Luisa Bisaglia and Margherita Gerolimetto</i>	
Microbiome Enterotype Detection via a Latent Variable Allocation Model	96
<i>Alice Giampino, Roberto Ascari, and Sonia Migliorati</i>	
Model-Based Clustering of Spatial Time Series Through the BayesMix library	102
<i>Matteo Gianella and Alessandra Guglielmi</i>	
Statistical Methods for Evaluating the Environmental Sustainability	108
<i>Andrea Gilardi and Francesca Ieva</i>	
Hierarchical Clustering for Three-Way Data	113
<i>Paolo Giordani, Susanna Levantesi, Andrea Nigri, and Donatella Vicari</i>	
Insights into Regression-Based Cross-Temporal Forecast Reconciliation	119
<i>Daniele Girolimetto and Tommaso Di Fonzo</i>	
Explainable Decision Tree Ensembles	126
<i>Agostino Gnasso, Massimo Aria, Carmela Iorio, and Marjolein Fokkema</i>	
The Study of the Empowerment in Two African Countries Using the Bayesian Networks	130
<i>Silvia Golia</i>	
Inverse Stereographic Gaussian DAG Models	136
<i>Anna Gottard and Agnese Panzera</i>	
Addressing Self-selection in Long Covid Studies: A Methodological Framework	142
<i>Konstantina Gourgoura, Elena Stanghellini, and Marco Doretti</i>	

An R-Shiny Application for Statistical Matching	148
<i>Fedele Greco, Lucia Guastadisegni, and Carlo Trivisano</i>	
On the Use of Simulations in the Construction of Composite Indexes	154
<i>Giulio Grossi and Emilia Rocco</i>	
Synthetic Control Method for Clinical Trials and Precision Medicine	160
<i>Veronica Ballerini and Giulio Grossi</i>	
Functional Clustering with Imbalanced Groups	166
<i>Catherine Higgins and Michelle Carey</i>	
A Triad Census-Based Synthetic Indicator to Summarise Cooperation in Football	171
<i>Riccardo Ievoli, Lucio Palazzo, Roberto Rondinelli, and Giancarlo Ragozini</i>	
Transition Towards Environmental Sustainability: Evaluating the Role of Green Investment, Environmental Policy, and Financial Development on the Load Capacity Factor in G-7 Countries	177
<i>Aamir Javed and Agnese Rapposelli</i>	
Offshore Wind Farm Development in Italy	183
<i>Feroz Khan and Agnese Rapposelli</i>	
Joint Complete Records Occurrence over Farlie-Gumbel-Morgenstern Sequences	188
<i>Amir Khorrami Chokami and Silvia Montagna</i>	
Deconvolution of the Electrical Load Curve Using Blind Source Separation	194
<i>Guillaume Koechlin, Filippo Bovera, and Piercesare Secchi</i>	
Forecasting Illiquidity Tail Risk and Its Economic Determinants	200
<i>Demetrio Lacava and Paolo Santucci de Magistris</i>	
Statistical Modelling of Population Gaps in Mortality: The Skellam Distribution to Model Differences in Death Counts	206
<i>Giacomo Lanfuiti Baldi and Andrea Nigri</i>	
The Challenge of Microdata Dissemination. A Methodology for a Public Use File Version of the WHIP Database	212
<i>Roberto Leombruni, Consuelo Rubina Nava, and Marina Zanatta</i>	

A Bootstrap Approach to Account for Model Selection in Multiple Systems Estimation 218
Lax Chan and Bernard W. Silverman

COVID-19 and Uncertainty in Italy. Exposure to High Mortality and Fertility, the Case Study on the City of Bergamo 223
Patrizio Lodetti and Victoria San Juan Bernuy

Enhancing Graphical Interpretability for Functional Random Survival Forest 230
Giuseppe Loffredo, Elvira Romano, and Fabrizio Maturo

Uncertainty in Sampling Designs for Non-probability Samples 234
Pier Luigi Conti and Daniela Marella

Model-Based Clustering of Nested Lifetime Data: Profiling COVID-19 Heart Failure Patients 240
Chiara Masci, Andrea Cappozzo, Francesca Ieva, Olivia Leoni, Marco Forlani, Barbara Antonelli, and Anna Maria Paganoni

Random Effects in Unconditional Quantile Regressions 246
Luca Merlo, Nicola Salvati, Enrico Fabrizi, Paolo Frumento, and Lea Petrella

Teachers’ Evaluations and Ethnic Penalty on Academic Achievement in Italy 251
Eleonora Miaci, Giulia Berta, Martina Sciarabba, and Nazareno Panichella

Improving Interpretability Using Surrogate Models: An Ecological Case Study 257
Gianni Gilioli, Manlio Migliorati, and Anna Simonetto

Statistical Analysis of Peer Grading: A Latent Variable Approach 263
Giuseppe Mignemi, Yunxiao Chen, and Irini Moustaki

Machine Learning and Air Pollution: An Empirical Application for Northern Italy 269
Angelo Mimmo and Daniele Toninelli

Big Data Subsampling: A Review 276
Rosa Arboretti, Marta Disegna, and Alberto Molena

Survey on Urban Environmental Data: Data Validation 282
Domenico Adamo, Gianpiero Bianchi, and Lucia Mongelli

Detecting Circadian Gene Expressions via Bayesian Analysis: An Application to the <i>Arabidopsis Thaliana</i> Dataset	288
<i>Silvia Montagna, Amir Khorrami Chokami, and Surya T. Tokdar</i>	
The Benefit of Social Capital Resources for Contrasting Ethnic Bullying at School: A Multilevel Study	294
<i>Maria Francesca Morabito, Serena Verbena, Valentina Tocchioni, and Benedetta E. Palladino</i>	
Computing Multidimensional Composite Indicators for Small Areas in Presence of Missing Variables: A Data Integration Approach	300
<i>Angelo Moretti, Alejandra Arias-Salazar, and Natalia Rojas-Perilla</i>	
A Tool for Identifying Drivers Impacting Students' Career Aspirations	306
<i>Flaminia Musella, Brunella Botte, and Giada Marinensi</i>	
Changes in Seasonal Concentration of Tourist Flows: The Role of Tourism Market Segments	312
<i>Mauro Mussini</i>	
Evaluating Volleyball Setters Using Product Partition Models	318
<i>Martina Narcisi, Garritt L. Page, and Gilbert W. Fellingham</i>	
Combining Price Indexes and Fuzzy Clustering to Measure Satisfaction in Niche Tourism Destinations	324
<i>Consuelo Rubina Nava, Tiziana Ciano, and Massimiliano Ferrara</i>	
The Efficacy of Social Distancing Measures in Controlling the Spread of SARS-CoV-2: A Retrospective Study in Italy	330
<i>Giorgia Bazzurri, Maura Mezzetti, and Ilia Negri</i>	
A Semi-parametric Spatial Model for Zero Inflated Weibull Distributions with Application to Extreme Rainfall Events	337
<i>Paolo Onorati and Antonio Canale</i>	
Has Immigration Contributed to Population Growth in Spain? A Regional Comparative Analysis	342
<i>G. Orfao, J. García-Gómez, A. Del Rey, and E. Parrado</i>	
A Fuzzy Clustering Approach to Detect the Number of States in Markov Switching Models	349
<i>Edoardo Otranto and Luca Scaffidi Domianello</i>	
Trimmed Symmetric Linear Regression	355
<i>Simona Pacillo, Luca Greco, and George Luta</i>	

Analysis of Complex Spatio-Temporal Neuroimaging Signals by Functional Principal Component Analysis	361
<i>Alessandro Palummo, Eleonora Arnone, Letizia Clementi, and Laura M. Sangalli</i>	
Asthma Prediction Tool in Preschool Children: Evidence from Predictive Performance Comparison by Using Innovative Statistical Approach	366
<i>Alessandra Pandolfo, Gianluca Sottile, Velia Malizia, Omar Shatarat, Valentina Lazzara, Vito Muggeo, Giovanni Viegì, and Stefaina La Grutta</i>	
The Integration Paths of Refugees and Asylum Seekers in Italy	373
<i>Camilla Pangallo, Cinzia Conti, Fabio Massimo Rottino, and Oliviero Casacchia</i>	
Species Distribution Modeling Approach for Biased Citizen Science Data	379
<i>Greta Panunzi, Jafet Belmont, Janine Illian, and Sara Martino</i>	
A Nonparametric Approach to Model Event-Data on Linear Networks	386
<i>Simone Panzeri, Aldo Clemente, Eleonora Arnone, and Laura M. Sangalli</i>	
Press Coverage of the Economy: Construction of Coverage and Sentiment Indices	392
<i>Paolomaria Fabrizio, Patrizio Lodetti, Raffaele Guetto, and Daniele Vignoli</i>	
“Stay Tuned!”: Validating a Questionnaire on Perceptions of University Students in Staying Informed with Current Events	398
<i>Anna Maria Parroco, Giulia Bagnasco, Morena Cucinella, Santina Pampillonia, and Micaela Arcaio</i>	
Ordinal Discriminative Dimensionality Reduction of Functional Profiles	404
<i>Giulia Patanè, Federica Nicolussi, Alexander Krauth, Günter Gauglitz, Bianca Maria Colosimo, Luca Dede’, and Alessandra Menafoglio</i>	
Reurbanize Small Towns by International Migration: A Possible Trade off a Possible Trade off a Possible Trade off	410
<i>Michela C. Pellicani, Antonella Bianchino, Monica Carbonara, Antonella Rotondo, Roberto A. Palumbo, and Massimo Iaquinta</i>	
Network Analysis: An Application to EU Geographical Indications	416
<i>Alfonso Peluso, Antonella Bianchino, and Agata Maria Madia Carucci</i>	
Immigrants’ Children and Socio-economic Achievement: An Analysis Across Western Europe	422
<i>Giorgio Piccitto</i>	

Unlocking Urban Insights: Leveraging Open Data for Governance and Research	426
<i>Giorgio Piccitto, Domingo Scisci, and Sonia Stefanizzi</i>	
Tourism Sustainability: A Statistical Study in Friuli Venezia Giulia Municipalities	429
<i>Gianfranco Piscopo, Gabriella Schoier, Massimiliano Giacalone, and Maria Longobardi</i>	
Forecasting Geneva's Donors and Their Charitable Deductions	435
<i>Marta Pittavino and Giedre Lideikyte-Huber</i>	
Bayesian Meta-learning Approach for Feasible Large Spatial Analysis	442
<i>Luca Presicce and Sudipto Banerjee</i>	
Social Inequality and Student Mobility from Southern Italian Regions	449
<i>Andrea Priulla, Eleonora Miaci, and Nazareno Panichella</i>	
A Bayesian Approach to Clustering Using the Proper Bayesian Bootstrap	455
<i>Federico Maria Quetti, Elena Ballante, and Silvia Figini</i>	
Analysis of Higher Education Dropouts Dynamics Through Functional Decomposition of Recurrent Events on Time-to-Event Processes	461
<i>Alessandra Ragni, Chiara Masci, and Anna Maria Paganoni</i>	
The PublicWorksFinanceIT R Package: Retrieve and Visualize Public Works Data	467
<i>Lorena Ricciotti</i>	
Oracle Priors in Bayesian Inference: a Frequentist View	473
<i>Stefano Rizzelli</i>	
Leveraging Pre-trained CNNs Feature Extractors for Classifying Genus <i>Oliva Bruguère</i> , 1789	479
<i>Luca Romagnoli, Luigi Ippoliti, Luigi Di Carlo, and Cesare Brizio</i>	
Explaining Item-Nonresponse in Open Questions with Requests for Voice Responses	483
<i>Camilla Salvatore and Jan Karem Höhne</i>	
IoT and AI for the Analysis of Emotional Intelligence	490
<i>Vito Santarcangelo, Saverio Gianluca Crisafulli, Giuseppe Oddo, Gianfranco Piscopo, and Massimiliano Giacalone</i>	

Leaders and Parties: How Do They Speak Online? 495
*Francesco Santelli, Alessia Forciniti, Valeria PolICASTRO,
and Giancarlo Ragozini*

Meta-analytic Metagenomic Analysis of the Microbiome in Gastric Cancer 501
*Gloria Santoro, Elena Locatelli, Flavio Tirelli, Annamaria Agnes,
Domenico D’Ugo, Roberto Persiani, Alberto Biondi, and Laura Lorenzon*

Estimating a Quantile-Based Income Inequality Indicator on Complex
Sampling Data 506
Silvia Scarpa, Maria Rosaria Ferrante, and Stefan Sperlich

Exploring New Perspectives on the Ukrainian-Russian Conflict Using
Text Mining Techniques: An Updated Analysis 511
Leonardo Scarso, Marco Novelli, and Francesco Saverio Violante

Competing Sets of Predictors in an Authorship Attribution Task: Most
Frequent Words, Large Language Models and Correspondence Analysis 517
Andrea Sciandra and Arjuna Tuzzi

Clustering Models for Time Series Analysis in Confocal Microscopy 524
Ivan Sciascia, Andrea Crosino, and Andrea Genre

Minimax Rates for Wasserstein Deconvolution of Regular Distributions
with Known Ordinary Smooth Errors 528
Catia Scricciolo

Capturing Uncertainty in Gaussian Mixture-Based Entropy Estimation
via Weighted Likelihood Bootstrap 534
Luca Scrucca

A New Approach for Text Bi-Clustering 540
Ilaria Bombelli, Domenica Fioredistella Iezzi, and Emiliano Seri

Language of Bankruptcy: Analysis of Word Sequences and Word Contexts 546
Andrea Simonetti and Rodolfo Damiano

Modelling Framework to Evaluate the Phenological Development of Vine
and Potential Impact of Frost: The Case of Prosecco DOC 552
Anna Simonetto, Eugenio Brentari, and Alfonso Piscitelli

A New Algorithm for the Imputation of Missing Functional Data 556
Marco Stefanucci

Exploring Disease Prevalence in Italy: A Web Application Utilizing Sample-Based Predictions from a Bayesian Spatio-Temporal Model with External Validation	563
<i>Mattia Stival, Lorenzo Schiavon, Gaia Bertarelli, and Stefano Campostrini</i>	
Dependent Infinite Latent Feature Models for Tree Occurrence Data	569
<i>Federica Stolf</i>	
War and the Military Service in Post-Communist Bloc: Which Impact on Marriage and Divorce?	574
<i>Sara Tafuro</i>	
Bayesian Inference for Discrete Spatio-Temporal Hidden Markov Models	581
<i>Daniele Tancini, Riccardo Rastelli, and Francesco Bartolucci</i>	
Modelling Three-Dimensional Point Patterns	588
<i>Marco Tarantino, Nicoletta D'Angelo, and Giada Adelfio</i>	
Presmoothing Methodology for Instrumental Variable Estimation of the Proportional Odds Model Under Random Right Censoring	595
<i>Lorenzo Tedesco</i>	
Measuring Political Cleavages Through the Decomposition of Diversity	601
<i>Giulio Giacomo Cantone and Venera Tomaselli</i>	
The Ripple Effect: How Pre-surgery Information Shapes Patient Satisfaction in Breast Cancer Surgery	608
<i>Costanza Tortú, Chiara Seghieri, and Francesca Ferré</i>	
Employment Trajectories and Fertility in Italy	613
<i>Marco Tosi, Fausta Ongaro, and Silvia Meggiolaro</i>	
Addressing Misclassification in the Assessment of Self-rated Health: Insights from the Elderly Population Italy	618
<i>Eleonora Trappolini, Maria Felice Arezzo, and Giuseppina Guagnano</i>	
Changes in Parenthood Expectancy Over Time Under an Age-Perspective	624
<i>Alessandra Trimarchi and Erich Striessnig</i>	
Bayesian Latent Factor Model for Multi-target Inference	630
<i>Bruno Ursino, Manuela Antonelli, Beatrice Cantoni, Ilenia Epifani, and Francesco Trovò</i>	

Clustering Multivariate Rating Data Within the CUB Framework 637
Matteo Ventura, Julien Jacques, and Paola Zuccolotto

Fairness-Adaptive Classification Trees 643
Anna Gottard, Sabrina Giordano, and Vanessa Verrina

An Ensemble Method for Disease Surveillance 649
Anna Vesely, Giulia Roli, Michele Scagliarini, and Rossella Miglio

Bayesian Inference for Compartmental Models with Intractable Likelihood
Function 655
Cecilia Viscardi, Alessio Lachi, and Michela Baccini

Author Index 661



Spatial Ranking Analysis on Quality of Services Using Some Techniques to Improved Fuzzy c-Means Algorithm

Najada Firza¹(✉), Laura Antonucci², and Dante Mazzitelli¹

¹ Department of Economics and Finance, University of Bari “Aldo Moro”, Largo Abbazia Santa Scolastica, 53, Bari, Italy
najada.firza@uniba.it

² Department of Clinical and Experimental Medicine, University of Foggia, Via L. Pinto, Foggia, Italy

Abstract. The Fair and Sustainable Wellbeing System (BES) with the related data and indicators proposed by the Institute of Official Statistics in Italy is closely linked to the Italian territory. Through the BES indicators monitored by ISTAT it is possible to trace the state of health of the territory. In this work we took into account all the BES indicators belonging to the “Quality of services” domain. We focused on this aspect because we believe that the quality of services is a very relevant domain for understanding the degree of sustainability of the territory with respect to the needs of its inhabitants. We started from a spatial cluster analysis trying to optimize the analysis by opting for a comparison of Euclidean and Mahalanobis dissimilarity matrix methods. Subsequently we built Key Performance Indicators through the Principal Components Analysis and finally a territorial ranking was drawn up to analyze and evaluate the quality of services in Italy.

Keywords: Fuzzy c-mean algorithm · PCA · Mahalanobis distance · Euclidean distance

1 Introduction

The UN Agenda 2030 emphasizes the necessity of adopting a fresh approach to development that respects the Earth’s resources while promoting an inclusive economy that reduces inequality. Access to comprehensive data measuring well-being and its sustainability, rather than solely relying on GDP, is essential for shaping new paradigms centred around well-being (Costanza et al., 2014).

According to Maggino (2017), BES serves as a system that portrays the well-being situation in various aspects within Italy and its regions.

It is important to note that the BES indicators follow an Italian approach and provide a wider range of measures for specific phenomena.

The BES system is also important to ensure uniform Italian policies at European level. This approach will allow for the creation of a global system that evaluates Italian well-being by considering different aspects and integrations.

2 Material and Methods

2.1 Data Collection

In this work we collected data from the ISTAT public database in the time window 2004–2022. We considered 16 indicators of “Quality of services” BES domain (<https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level>, Dataset BES at local level 2023 edition). Data were normalized with the Range 0–1 method (Table 1).

Table 1. BES domain Indicators for Quality of Services at Regional Italian level.

Quality of services	1	Beds in the residential social-healthcare and social-welfare facilities	Beds in the public or private social-healthcare and social-welfare facilities that provide residential services (assisted hospitality with overnight stay) of a social welfare and/or social-healthcare type to people in need per 10,000 inhabitants	Istat - Residential health and social care facilities
	2	Integrated home assistance service	Percentage of people aged 65 and over who benefited from integrated home assistance service	Istat - Processing of data from Ministry of Health
	3	Composite index of service accessibility	Percentage of households who find very difficult to reach some basic services (pharmacy, emergency room, post office, police, carabinieri, municipal offices, crèches, nursery, primary and secondary school, market and supermarket). The indicator is a three-year average	Istat - Survey on Aspects of daily life
	4	Irregularities in water supply	Percentage of households who report irregularities in water supply	Istat - Survey on Aspects of daily life

(continued)

Table 1. *(continued)*

5	Irregularities in electric power distribution	Frequency of accidental long lasting electric power cuts (cuts without notice longer than 3 min) (average number per consumer)	Istat - Processing of data from the Italian Regulatory Authority for Energy, Networks and Environment (Arera)
6	Seat-Km of public transport networks	Seat-Km of public transport networks per capita in provincial capital Municipalities	Istat - Survey on urban environmental data
7	Satisfaction with public transport services	Percentage of people aged 14 years and over, regular users of public transport services, who rate their experience of these services as positive (grade of 8/10 or higher). Regular users are those reporting to use public transport (train or bus/trolleybus/tram, urban or suburban) more than once a week	Istat - Survey on Aspects of daily life
8	Frequent users of public transport	Percentage of population aged 14 and over who use public transport several times a week (bus, trolley bus, tram within their own municipality; bus or coach connecting different municipalities; train)	Istat - Survey on Aspects of daily life
9	Overall Fixed Very High Capacity Network (VHCN) coverage	Percentage of households which are resident in an area served by a very high capacity network (FTTH)	Istat - Processing of data from Agcom

(continued)

Table 1. *(continued)*

10	Separate collection service for municipal waste	Percentage of resident population in municipalities with separate collection greater than or equal to 65%	Istat - Processing of data from ISPRA
11	Hospital beds in high-care wards	Hospital beds in high care wards in ordinary inpatient care in public and private health care institutions per 10,000 inhabitants	Istat - Processing of data from Ministry of Health
12	Hospital patient emigration to a different region	Percentage of hospital discharges carried out in regions other than that of residence on the total discharges of residents in the region. Data refer only to inpatient discharges for “acute” care (excluding hospitalizations of “spinal unit”, “functional rehabilitation”, “neuro-rehabilitation”, “long-term care”)	Istat - Processing of data from Hospital Discharges of the Ministry of Health
13	Unmet need for medical examination	Percentage of the population reporting unmet needs for medical care in the last 12 months (specialist visits - excluding dentist visits - or diagnostic examination) due to one of the following reasons: financial reasons; inconvenience (distant structure, lack of transportation, inconvenient hours); long waiting list; Covid (since 2020)	Istat - Survey on Aspects of daily life

(continued)

Table 1. (continued)

14	General practitioners with a number of patients above the maximum threshold	Percentage of general practitioners with a number of patients above the maximum threshold of 1500 patients defined by the general practitioners' contract	Istat - Processing of data from Ministry of Health
15	Physicians	Practising physicians per 1,000 inhabitants	IQVIA ITALIA
16	Nurses and midwives	Practicing nurses and midwives per 1,000 inhabitants	Co.Ge.A.P.S

Source: ISTAT BES Indicators

2.2 The Fuzzy c-Means

The method used for grouping is fuzzy c-means (FCM). A partition clustering algorithm related to the k-means algorithm [1]. This fuzzy c-means clustering algorithm is capable of partitioning objects simultaneously taking into account several dissimilarity matrices:

$$J = \sum_{i=1}^M \sum_{j=1}^N u_{ij}^p \|x_i - c_j\|^2$$

The algorithm is designed to provide a fuzzy partition and profile for each fuzzy cluster, as well as learn the specific weight of relevance for each dissimilarity matrix all this to optimize an objective function. It assigns by randomly a weight of membership u_{ij} to each element x_i which belongs to each cluster j that we want to find. By means of an iterative process, an objective function composed by sum of the distances of each point from each cluster centre is minimized, c_i is the centroid of cluster J ; M is the number of instances in the dataset; N is the desired number of clusters imposed by the researcher; p is the fuzzifier parameter, a hyper-parameter that controls the “fuzzy degree” of the cluster. Firstly, we used the Euclidean distance metric to compute the distances between the cluster centres and each observation in the sample (Bezdek et al. 1974).

In second step, we propose to utilize the Mahalanobis distance to define the dissimilarity measure in this way reflecting the covariance of the cluster:

$$d_{ij} = (x_i - \mu_j)^T \sum_j^{-1} (x_i - \mu_j)$$

Where $\sum_j$ the covariance matrix of the j th cluster.

To have more robust analyses we used two different validation cluster methods:

- Silhouette index;
- Xie Beni index.

2.3 The Key Performance Indicators

We built a synthetic indicator to rank the Italian provinces. These indicators, called Key Performance Indicators (KPI), are based on both Principal Component Analysis (PCA) and Factor Analysis (FA).

The KPI is defined as:

$$KPI = \frac{\sum_{i=1}^N c_i v_i}{\sum_{i=1}^n c_i}$$

3 Results and Discussion

In this work we have developed an analysis framework to study 16 indicators belonging to the **Quality of Services** BES domain at the Italian regional level for a time interval of 19 years.

Italy is a nation known for its long-standing emphasis on various territorial disparities, stemming from what is commonly referred to as the NORTH-SOUTH divide (Alaimo & Maggino, 2020). Our analyzes confirm this gap (Table 2).

The analyzes processed with R and SPSS software. To obtain results with the PCA method, we used the procedure that is known as “backward elimination”. It consists of selecting all components with sufficient eigenvalue (in this case $> 1,05$) and deleting the items with low communality, starting from the lowest value.

The procedure stopped when 4 principal components are identified, explaining over 79% of the total variance.

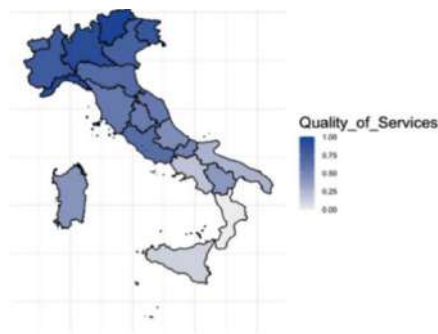
In order to better interpret such results, a Varimax orthogonal rotation was applied to those principal components. This approach is used to identify the existence of significant relationships between the indexes and the latent variables which can better explain the phenomenon (Kaiser, 1958).

Comparisons of all outcomes can be made with the KPI ranking and the ISTAT 2023 BES Report or ISTAT 2023 SDGs Report. The ongoing pandemic has resulted in a leveling of disparities in terms of access to healthcare services, particularly with a notable increase in the northwest, northeast, and central regions. According to the ISTAT SDGs Report in 2021, 9.4% of households report irregular water supply in their homes (compared to 8.9% in 2020).

These irregularities vary across regions, with a higher prevalence in the South, affecting 1.5 million families (63.9% of the total), notably in Sicily (29.0%) and Calabria (28.8%). In contrast, the figures are significantly lower in the northwest and northeast regions (3.1% and 3.5% respectively), while in the central region, less than one in ten families report irregularities in service. Italy exhibits pronounced fragmentation in management, particularly in areas where integrated water supply systems have not been fully implemented, such as Calabria, Molise, the autonomous provinces of Trento and Bolzano, Sicily, and Valle d’Aosta. This fragmentation may impede the widespread coverage of the territory by water security plans, as highlighted in the SDGs ISTAT Report 2022 (Fig. 1).

Table 2. Ranking of the Italian Region in the Quality of Services BES domain.

Position	Quality of Services
1	Trentino A. A.
2	Liguria
3	Lombardy
4	Friuli V. G.
5	Piedmont
6	Veneto
7	Aosta Valley
8	Emilia Rom
9	Lazio
10	Tuscany
11	Marche
12	Umbria
13	Molise
14	Abruzzo
15	Sardinia
16	Basilicata
17	Apulia
18	Campania
19	Calabria
20	Sicily

**Fig. 1.** Geographical distribution of KPI for the Quality of Services BES domain.

4 Conclusions

In this article we analyzed Italy at a regional level. We focused on 16 indices from the BES Quality of Services domain. Our analyzes underline the strong existence of the North-South divide for the Service Quality domain.

Key performance indicators (KPIs) are a valuable tool for measuring and monitoring progress towards these goals. They facilitate ongoing monitoring and evaluation of progress, highlight areas for improvement and guide targeted interventions.

References

1. D'Urso, P., Alaimo, L.S., De Giovanni, L., et al.: Well-being in the italian regions over time. *Soc. In dic Res* **161**, 599–627 (2022). <https://doi.org/10.1007/s11205-020-02384-x>
2. Alaimo, L.S.: Sustainable development and national differences: an european cross-national analysis of economic sustainability. *RIEDS-Rivista Italiana di Economia, Demografia e Statistica - The Italian Review of Economics, Demography and Statistics* **72**(3), 101–123 (2018)
3. Alaimo, L.S., Maggino, F.: Sustainable development goals indicators at territorial level: conceptual and methodological issues—the italian perspective. *Soc. Indic. Res.* **147**(2), 383–419 (2020). <https://doi.org/10.1007/s11205-019-02162-4>
4. Costanza, R., Hart, M., Posner, S., Talberth, J.: “Beyond GDP: The Need for New Measures of Progress.” Pardee Paper No. 4, Boston: Pardee Center for the Study of the Longer-Range Future (2009)
5. Costanza, R., et al.: Development: time to leave GDP behind. *Nature* **505**(7483), 283–285 (2014). <https://doi.org/10.1038/505283a>
6. Kaiser, H.F.: The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200 (1958)
7. Maggino, F.: Developing indicators and managing the complexity. In: Maggino, F. (ed.) *Complexity in Society: From Indicators Construction to their Synthesis*, pp. 87–114. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-60595-1_4
8. https://www.istat.it/it/files/2023/05/2022-SDGS-Report_Inglese.pdf
9. <https://www.istat.it/it/files/2023/10/BES-2022-Report.pdf>



A Topology-Based Algorithm for the Isomorphism Check of 2-Level Orthogonal Arrays

Roberto Fontana¹ and Marco Guerra^{2(✉)}

¹ Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli
Abruzzi 24, 10129 Turin, Italy
`roberto.fontana@polito.it`

² Institut Fourier, Université Grenoble-Alpes, CNRS 100, Rue des Mathématiques,
38610 Gières, France
`marco.guerra@univ-grenoble-alpes.fr`

Abstract. We introduce a construction and an algorithm, both based on Topological Data Analysis (TDA), to tackle the problem of the isomorphism check of Orthogonal Arrays (OAs). Specifically, we associate to any binary OA a persistence diagram, one of the main tools in TDA, and explore how the Wasserstein distance between persistence diagrams can be used to inform whether two designs are isomorphic.

Keywords: Topological Data Analysis · Orthogonal Arrays · Isomorphism Problem

1 Introduction

Factorial designs are commonly used in various fields of application, including engineering and agriculture. The isomorphism check determines whether two designs are equivalent. For qualitative factors, two designs are combinatorially isomorphic if one design can be obtained from the other by (i) reordering the runs, (ii) relabeling the factors and/or (iii) switching the levels of one or more factors. For quantitative factors, two designs are geometrically isomorphic if one design can be obtained from the other by (i), (ii), and/or reversing the level order of one or more factors. For 2-level designs, combinatorial and geometric isomorphism are equivalent. The isomorphism check has been extensively studied in the statistical literature (see [1, 3, 5–7, 9]), and is a known hard computational problem.

In this work we focus on the isomorphism check for *binary* factorial designs, specifically, we work on the subclass of Orthogonal Arrays of strength 2. However, our method can be applied to OAs of any strength or, more generally, to any type of 2-level designs. We consider OAs because the set of non-isomorphic OAs for different number of factors is easily available (e.g. we used the Orthogonal Array package, [8]), so the performance of our algorithm could be easily assessed.

We introduce a simple construction that associates to any d -variate Bernoulli random variable a *persistence diagram*, a key tool in TDA. By doing so, and by interpreting an OA as a multivariate Bernoulli, we make use of a notion of Wasserstein distance between persistence diagrams to inform the isomorphism check problem, via a statistical procedure. Notice the 2-level hypothesis is required for the construction to work. We report good performance of our algorithm in the cases we have explored.

2 The Isomorphism Check Problem

2.1 Factorial Designs and Orthogonal Arrays

Given d binary factors $A_1, \dots, A_d$, whose levels are coded as $\{0, 1\}$, we represent a fraction $\mathcal{F}$ of the full factorial design $\mathcal{D} = \{\alpha \in \{0, 1\}^d\}$ using the counting vector $v_{\mathcal{F}}$. Each component v_{α} of the counting vector $v_{\mathcal{F}}$ is the number of times that the point α appears in $\mathcal{F}$, $v_{\mathcal{F}} = (v_{\alpha}, \alpha \in \{0, 1\}^d)$. It follows that $v_{\alpha} \geq 0, \alpha \in \{0, 1\}^d$. We order the set $\{\alpha \in \{0, 1\}^d\}$ lexicographically.

Pmf and Moments. We observe that each vector $v_{\mathcal{F}}$ determines $p_{\mathcal{F}}$, the probability mass function (pmf) of a d -variate Bernoulli random variable $X = (X_1, \dots, X_d)$ simply by dividing each component of $v_{\mathcal{F}}$ by N , the number of points contained in $\mathcal{F}$: $p_{\mathcal{F}} = \frac{1}{N}v_{\mathcal{F}}$. We also consider $\mu_{\mathcal{F}}$, the vector of moments of $X \sim p_{\mathcal{F}}$, $\mu_{\mathcal{F}} = (\mu_{\alpha}, \alpha \in \{0, 1\}^d)$, where $\mu_{\alpha} = E[X^{\alpha}] \equiv E[X_1^{\alpha_1} \dots X_d^{\alpha_d}]$.

For example, for $d = 2$, the counting vector $v_{\mathcal{F}}$ of $\mathcal{F} = \{(0, 0), (0, 0), (0, 1), (1, 0), (1, 1), (1, 1)\}$ is $v_{\mathcal{F}} = (2, 1, 1, 2)$, the corresponding pmf is $p_{\mathcal{F}} = (1/3, 1/6, 1/6, 1/3)$, and $\mu_{\mathcal{F}} = (1, 1/2, 1/2, 1/3)$.

OAs. An OA of strength t on N runs, with d binary factors, $1 \leq t \leq d$, is a fraction $\mathcal{F}$ of N possibly repeating points of $\mathcal{D} = \{0, 1\}^d$ such that every projection $\mathcal{F}_t$ of $\mathcal{F}$ on a subset of t factors is a full factorial design with t factors and each run of $\mathcal{F}_t$ appears the same number of times, $N/2^t$. The fraction $\mathcal{F}$ above is an OA of strength $t = 1$ and size $N = 6$.

Isomorphism. Two fractions $\mathcal{F}_1$ and $\mathcal{F}_2$ of $\mathcal{D}$ are said to be isomorphic if one is obtained from the other by a combination of (i) reordering the runs, (ii) relabeling any of the factors, (iii) switching the levels on any factor.

For example the OA $\mathcal{F}_1 = \{(0, 1), (0, 1), (0, 0), (1, 1), (1, 0), (1, 0)\}$ is isomorphic to $\mathcal{F}$ above; it has been obtained from $\mathcal{F}$ by switching the levels on factor A_2 .

2.2 Topological Data Analysis

Topological data analysis (TDA) is a recent subject of mathematics that leverages classical tools from algebraic topology, such as homology theory, to study data. In the last 25 years it has found wide application, partly thanks to its sound theoretical grounding (see [2] for a review). The main tool in TDA is called persistent homology (PH); intuitively, homology describes the *shape* of a topological object via algebraic invariants. Data, in a very broad sense, can be

given the structure of a family of topological spaces, each described by homology. Persistent homology, then, describes the *shape of data* by tracking the evolution of homology through this family of spaces associated with data.

A proper introduction of PH cannot fit here, therefore we refer to the bibliography ([4]) and only sketch the concepts we will use. PH uses the homology of simplicial complexes, which are discrete topological spaces based on standard building blocks called simplices, suitable for computation. Data in a metric space yields an *increasing family* of simplicial complexes, called a *filtration*, i.e. a family of simplicial complexes that are totally ordered by inclusion. Simplicial homology counts the *obstructions to connectivity*, that is, connected components in dimension 0, holes in dimension 1, voids in dimension 2, and so on. Tracking simplicial homology along a filtration yields persistent homology, that is the birth and death of connected components, holes, voids and so on. This information is summarized in the *persistence diagram*, which we denote by Dgm , a set of points in the plane, each referring to a homological feature (hole, void, etc.), whose first coordinate is its birth time and second coordinate its death time. Note features that persist until the end of the filtration have infinite second coordinate.

2.3 The Dgm of an OA

The full factorial design $\mathcal{D} = \{\alpha \in \{0, 1\}^d\}$ is a simplicial complex, namely the $d - 1$ simplex plus all its faces. Intuitively, introduce d abstract vertices: each α denotes which vertices are “switched on” (have a 1), and therefore defines a simplex, in particular a face of the $d - 1$ simplex. Given a Bernoulli pmf on $\mathcal{D}$, the moments vector μ associates a real number μ_α to each simplex α . We state here without proof that the sublevel sets of the function $1 - \mu_\alpha : \mathcal{D} \rightarrow \mathbb{R}$ define a filtration of simplicial complexes. Therefore, to each pmf we can associate a well-defined persistence diagram. By the above, we can do the same to any binary OA.

2.4 Wasserstein Distances

The Wasserstein distance is a distance between two probability measures over a common metric space. In this work we consider the Wasserstein distance $d_W^1(p_1, p_2)$ between two pmfs, p_1 and p_2 , that correspond to the fractions $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. The distance is computed by finding an optimal permutation π , that maps the two vectors as close as possible

$$d_W^1(p_1, p_2) := \inf_{\pi} \sum_{\alpha \in \{0, 1\}^d} |p_1(\alpha) - p_2(\pi(\alpha))|.$$

If (and only if) $d_W^1(p_1, p_2) = 0$, then the two pmfs are a permutation of one another. It follows that if $d_W^1(p_1, p_2) > 0$ then $\mathcal{F}_1$ and $\mathcal{F}_2$ are not isomorphic.

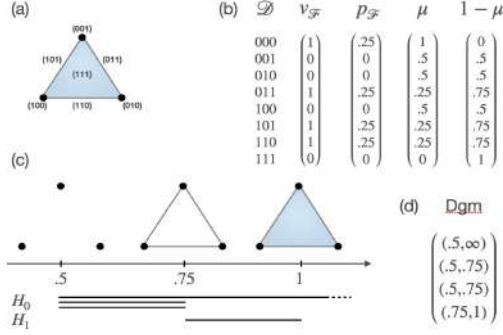


Fig. 1. (a) The simplices corresponding to the $\alpha \in \{0,1\}^3$. (b) A 4-point OA in $d = 3$, and its corresponding pmf, moment vector, and filtration $1 - \mu$. Notice the α 's are ordered lexicographically in the vectors. (c) The corresponding filtration made of sublevel sets of $1 - \mu$. The bars represent persistence of homological features (3 connected components and one loop). (d) The persistence diagram showing the pairs.

The Wasserstein distance between two persistence diagrams Dgm_1 and Dgm_2 corresponding to $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively, is obtained by finding a bijection between the finite points of the diagrams as points in $\mathbb{R}^2$, where the metric is the Euclidean one.

$$d_W^2(Dgm_1, Dgm_2) := \inf_{\eta: Dgm_1 \leftrightarrow Dgm_2} \left(\sum_{x \in Dgm_1} \|x - \eta(x)\|_2 \right)$$

If (and only if) $d_W^2(Dgm_1, Dgm_2) = 0$, then the two PD's are a permutation of one another. This implies the filtration vectors are a *pairwise* permutation of one another, i.e. a permutation that is compatible with the pairing of simplices given by the persistence pairs.

3 The Algorithm

We propose the following statistical algorithm to test if two OAs, $\mathcal{F}_1$ and $\mathcal{F}_2$ are isomorphic. We denote by p_i (Dgm_i) the pmf (the persistence diagram) corresponding to $\mathcal{F}_i$, $i = 1, 2$, respectively, and by T the maximum number of iterations.

Given $\mathcal{F}_1$ and $\mathcal{F}_2$ the corresponding pmf's p_1 and p_2 are computed. If the Wasserstein distance $d_W^1(p_1, p_2)$ between p_1 and p_2 is positive a permutation π such that $p_2 = \pi(p_1)$ does not exist and then $\mathcal{F}_1$ and $\mathcal{F}_2$ are not isomorphic. If $d_W^1(p_1, p_2) = 0$ a permutation π such that $p_2 = \pi(p_1)$ exists. We must decide if the permutation is a combination of (i) reordering the runs, (ii) relabeling any of the factors, (iii) switching the levels on any factor or not. To do that we compute $d_W^2(Dgm_1, Dgm_2)$, the Wasserstein distance between the persistence diagrams Dgm_1 and Dgm_2 corresponding to $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. If the distance is zero

the algorithm indicates that $\mathcal{F}_1$ and $\mathcal{F}_2$ are isomorphic. If the distance is greater than zero the algorithm considers T fractions $\mathcal{F}_2^{(1)}, \mathcal{F}_2^{(2)}, \dots, \mathcal{F}_2^{(T)}$ isomorphic to $\mathcal{F}_2$ and the corresponding persistence diagrams $Dgm_2^{(1)}, Dgm_2^{(2)}, \dots, Dgm_2^{(T)}$. If any of the distances $d_W^2(Dgm_1, Dgm_2^{(t)}), t = 1, \dots, T$ is zero than the algorithm indicates that $\mathcal{F}_1$ and $\mathcal{F}_2$ are isomorphic, otherwise $\mathcal{F}_1$ and $\mathcal{F}_2$ are considered as non-isomorphic.

```

Input  $p_1, p_2$ , and  $T$ 
Output: True/False, where True means that  $\mathcal{F}_1$  and  $\mathcal{F}_2$  are isomorphic.
if  $d_W^1(p_1, p_2) > 0$  then
    Status:=False;
else
    Status:=False
    for  $i = 1, \dots, T$  do
        Compute  $f_1, f_2$ 
        if  $d_W^1(f_1, f_2) == 0$  then
            Compute  $Dgm_1, Dgm_2$ 
            if  $d_W^2(Dgm_1, Dgm_2) == 0$  then
                Status:=True
            end if
             $p_2 \leftarrow$  Sample  $p_2'$  isomorphic to  $p_2$ 
        end if
    end for
end if
return Status;

```

4 Applications

We consider binary OAs for $d = 5$, size $N = 20$, and strength $t = 2$. There are 11 classes of non-isomorphic OAs for these parameters. We denote by $\mathcal{F}_0 \dots, \mathcal{F}_{10}$ the 11 non-isomorphic OAs that we have computed using the Orthogonal Array package, [8] and by $p_0 \dots, p_{10}$ the corresponding pmfs. We observe (see Fig. 2 left) that there are cases where the Wasserstein distance between pmfs fails to discriminate non-isomorphic arrays; for example the $d_W^1(p_3, p_6) = 0$. Conversely, in all cases the criterion based on the d_W^2 between persistence diagrams is able to correctly classify isomorphic/non-isomorphic OAs (see Fig. 2 right). We have also tested the algorithm in the cases $d = 3, 4$ ($N = 20, t = 2$), with equally correct results.

5 Concluding Remarks

To the best of our knowledge, this is the first time a TDA-based method has been applied to solving the isomorphism check for binary orthogonal arrays.

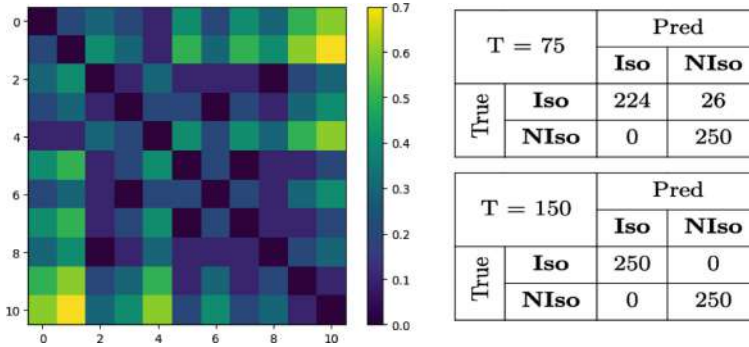


Fig. 2. Left, d_W^1 between pmfs of the 11 classes for $d = 5$, $N = 20$, $t = 2$. Right, the classification results of the algorithm for a sample of 500 OAs of the same parameters for $T = 75$ and $T = 150$. The second case achieves perfect classification. We sample from a set of 10,752 OAs.

For sufficiently large T , the algorithm was successful in all cases for $d = 3, 4, 5$, $N = 20$, $t = 2$ that we have considered. Our main goal is to prove that the statistical procedure is in fact exact at least in one direction:

Proposition 1 (Conjecture). *If $\mathcal{F}_1$ and $\mathcal{F}_2$ are not isomorphic, then*

$$d_W^2(Dgm_1, Dgm_2) > 0$$

References

1. Angelopoulos, P., Evangelaras, H., Koukouvinos, C., Lappas, E.: An effective step-down algorithm for the construction and the identification of nonisomorphic orthogonal arrays. *Metrika* **66**, 139–149 (2007)
2. Chazal, F., Michel, B.: An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Front. Artif. Intell.* **4** (2021). <https://doi.org/10.3389/frai.2021.667963>
3. Clark, J.B., Dean, A.M.: Equivalence of fractional factorial designs. *Statistica Sinica* 537–547 (2001)
4. Edelsbrunner, H., Harer, J.: *Computational Topology: An Introduction* (2010). https://doi.org/10.1007/978-3-540-33259-6_7
5. Lin, C.D., Sitter, R.: An isomorphism check for two-level fractional factorial designs. *J. Statist. Plan. Inference* **138**(4), 1085–1101 (2008)
6. Lin, C.Y., Cheng, S.W.: Isomorphism examination based on the count vector. *Statistica Sinica* 1253–1272 (2012)
7. Pang, F., Liu, M.Q.: Geometric isomorphism check for symmetric factorial designs. *J. Complex.* **27**(5), 441–448 (2011)
8. Schoen, E.D., Eendebak, P.T., Nguyen, M.V.: Complete enumeration of pure-level and mixed-level orthogonal arrays. *J. Comb. Des.* **18**(2), 123–140 (2010)
9. Wu, C., Xu, H.: Generalized minimum aberration for asymmetrical fractional factorial designs. *Ann. Stat.* **29**(2), 549–560 (2001)



The Revolution of AI in Healthcare Between Diagnosis and Treatment

Alessia Forciniti¹(✉) and Francesco Santelli²

¹ IULM University, 20143 Milan, Italy
alessia.forciniti@iulm.it

² University of Trieste, 34127 Trieste, Italy
fsantelli@units.it

Abstract. The complexity and volume of data in healthcare entail that artificial intelligence (AI) and associated technologies are becoming an essential component of life sciences. The scientific literature explores the benefits, contexts of application, ethical implications, and future developments. In order to 1) identify the topics characterizing the literature on AI in the clinical domain, 2) detect semantic categories, and 3) validate them, a methodological approach based on the combination of natural language processing and machine learning for classification was performed. Two main semantic categories were identified: diagnostics and treatment, which are used to manually annotate each document. Finally, we tested our semantic classification through machine learning. The findings suggest clear differences between the two categories, mainly based on AI-assisted meta-analyses and clinical decision support systems, with just a quota of scientific papers encompassing both semantic pillars. This proportion of documents is pivotal to changing the semantic classification.

Keywords: AI · NLP · embedded topic model · SVM

1 The Disruptive AI in Healthcare

The introduction of artificial intelligence (AI) and related technologies have contaminated the most diverse fields, spreading rapidly into business, industries, society, and life sciences.

Significant advances are emerging in the health sector [6], where different types of AI have been employed to contribute to more efficient management of administrative activities, to improve the pharmaceutical organisations, and for a cost reduction [5], to help in patient care and diagnosis [8], to improve inclination to care, to manage patients and medical resources. The approaches adopted include neural networks (for categorizing purposes, such as evaluating if a patient would get a specific illness), deep learning (for cancerous recognition of lesions in pictures), machine learning (mainly used as precision medicine, which predicts which treatment protocols are likely to be effective on a patient based

on numerous patient characteristics and the treatment environment). Also dominant is the natural language processing (NLP), widely used for statistical and semantic analysis on large corpora of clinical nature, or to transcribe interactions with patients or simulate conversational AI. Noteworthy are the robotic surgical technologies aimed at improving the performance in the operating room, through precise and less invasive incisions. With regard to information systems, AI is able to support working areas, corporate and integration processes through robotic process automation (RPA). In recent years, there have been an increasing number of scientific publications aimed at studying the advantages, technologies, ethical implications and future developments of AI in healthcare. In order to study the literature on AI advances in healthcare and identify the main transformative pillars, we analyzed a corpus of scientific publications from 2000 to 2023. An approach based on a combination of NLP and a supervised model with associated learning algorithms for classification have been performed. The research directions pursued are: a) to identify the topics that have characterized the literature on AI in the clinical domain; b) to determine representative semantic pillars; and c) to validate the semantic classes.

Section 2 presents data collection; Sect. 3 describes the methods; Sect. 4 shows results and conclusions.

2 Data

The data collection is composed of literature released until January 19th, 2024 from the bibliographic database Web of Science (WoS). For the data extraction, we used two different searches: for the first query, we used the keywords “artificial intelligence” and “health”, while for the second one, we used “artificial intelligence” and “clinical”. We have selected the years from 2000 to 2023, filtering only those in the English language and extending the query to every type of publication (paper, chapter, book, etc.). The result of the two extractions was integrated into a single dataset consisting of $N = 10,871$ abstracts and 72 metadata such as authors, titles, number of quotes, and more. From the data integration, the duplicates were removed, and the final corpus is composed of $N = 10,775$ documents with a prevalence of paper (94.46%). The first part of the 2000s is quite limited in the literature and shows exponential growth in recent years (Fig. 1).

The scientific literature in 2000 reflected a growing recognition of the potential of AI technologies in applications ranging from clinical decision support and medical imaging to data mining and genomics. Since 2017, there has been a significant growth of studies based on machine learning and deep learning for risk stratification, patient outcome prediction, and Clinical Decision Support Systems (CDSS). Many studies are focused on automated detection and classification of abnormalities in images (e.g., X-rays, MRIs, CT scans) or on improving information retrieval, decision support, and clinical workflow efficiency through NLP. Finally, the 2019–2023 slot saw the largest number of publications, mainly focused on the new frontiers of robotics and on ethical implications and future challenges.

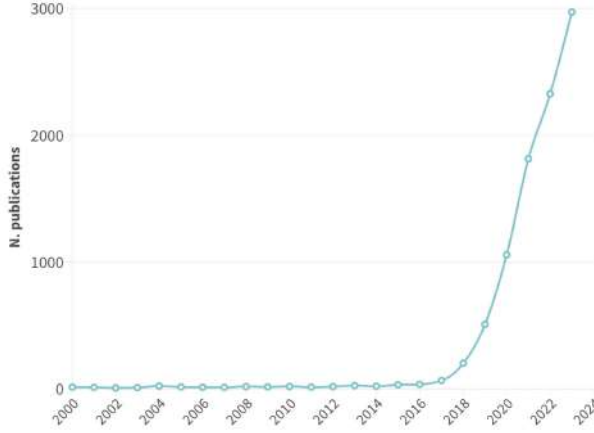


Fig. 1. Time plot of scientific literature from 2000 to 2023

3 Methods

3.1 Natural Language Processing (NLP)

To answer the first objective, we performed a machine learning model based on topic modeling in embedding spaces called ETM [4]. At first, we pre-processed the texts through tokenization, normalization, and lemmatization, removal of numbers, punctuation, English stop words (list ISO 639-1), and keywords used for data extraction. We used a TF-IDF weight to capture a greater discriminant power of the words. Finally, we performed the topics extraction through ETM, which is based on the discovery of semantic structures using the properties of latent Dirichlet allocation [2] and word embedding [1]. While LDA configures itself as a hierarchical probabilistic model, through which each topic is represented by a distribution of terms and each document by a mixture of the topics, in word embedding, the learning of a distributed representation means that the words with similar meanings are close in a lower-dimensional vector space. The ETM is more robust than the LDA when the corpora size increases [4]. To meet the second research direction, through the topics that emerged from ETM, we detected the main semantic pillars of AI applications. Based on these pillars, we manually annotated texts classified them and built a sub-corpus of annotated documents.

3.2 Support Vector Machine for Semantic Classification

To test our semantic annotation, the support vector machine (SVM) [3] was adopted. The model classifies mapping data as points in space, dividing categories with as large a gap as feasible. It determines the best way to separate hyperplanes. Specifically, in our study, we adopted a SVM with a linear Kernel, where the hyperplane that maximizes the margin between different classes in

the feature space is linear. The model operates without explicitly transforming the data into a higher-dimensional space but by finding the optimal separating hyperplane directly in the input space. The margin optimization solution is able to maximize the margin while minimizing classification errors. SVMs are useful for text and hypertext classification because they may considerably minimize the requirement for labeled training cases in both classic inductive and transductive situations [7], as well as semantic parsing.

4 Results and Conclusion

4.1 NLP

The corpus is characterized by 19,685 tokens and 10,131 types. The lexical diversity was measured through the Type-Token ratio (TTR), which shows that a good variety equals 51.46%. The average of words presenting just one time through the hapax is equal to 0.4, and the average length of sentences is 6.55, while the root mean square is 2.49. The feature extraction returned 7,575 terms and 98.94% sparsity of lexical table. Through the ETM, we have identified 43 different topics capable of outlining diverse application contexts, ranging from oncology to performance, from perioperative to prediction, from governance to algorithms, from empowerment to bias, from robotics to pathology and to more specialized clinical fields such as ophthalmology, endocrinology. There are 328 terms that characterize the 43 topics, extracted with the imposition of feature extraction of the top 50% terms with the highest TF-IDF. By imposing a hyperparameter for variational inference coded in dimension 100 for the learning, we obtained a good classification both on 70% of the training set and on 30% of the test set. Through the ETM, two main pillars were identified: one concerns the domain of diagnostics and the other concerns treatment. However, the topics and terms were not always clearly distinguishable between the diagnosis and treatment, but they sometimes overlapped to some extent. Based on these two semantic categories, we have labeled each document in the entire dataset. From the annotation, there were 3,397 documents related to the category “diagnosis” and 2,936 for the category “treatment”. A sub-corpus of these 6,333 documents was used for the SVM classification model.

4.2 SVM

The SVM model with linear Kernel performed the classification with 5,066 samples, 57 predictors, and the two classes “diagnosis” and “treatment”. The confusion matrix identifies 70.9% of true positive cases and 71.7% of true negative (Fig. 2). For cross-validating, the resampling returned a good accuracy of 0.71 as a proportion of correctly classified observations, with Tuning parameter ‘C’ (Cost) constant of 1. The precision measure equal to 0.71 means that the algorithm returned more relevant results than irrelevant ones with good performance. The recall equal to 0.77 shows most of the relevant results (whether or

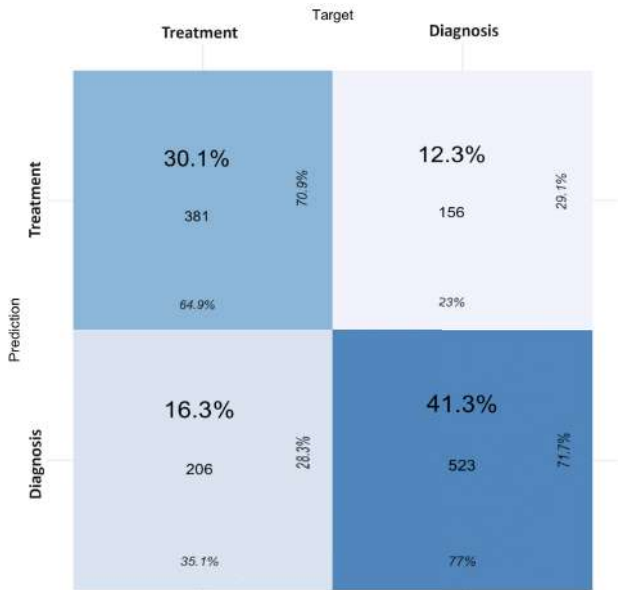


Fig. 2. Confusion Matrix (SVM classification) on entire corpus

not irrelevant ones are also returned). The error metric, which measures model performance by F1-score, confirms a good classification.

However, performance measurements and the percentage of false negatives and false positives suggest that we should take into consideration the limitation of this specific outcome of the ETM, in which we found the presence of topics belonging to both diagnosis and treatment at the same time. Therefore, we removed from the corpus all documents that contained lexical references classifiable in both semantic pillars, obtaining a corpus of 5,093 documents on which replicate the SVM analysis (Fig. 3).

In this case, the model used 4,075 samples and improved performance: we have reached 100% of the real positive and 88.5% of the true negative; the accuracy is 0.94, the precision is 0.88, the recall is 1, and the F1 score is 0.94.

The methodological approach applied in this study by combining NLP and classification models proved to be effective in document categorization. The joint use of topic models in embedding spaces has managed to identify the most characteristic semantic dimensions of the literature on AI in healthcare from 2000 to 2023. The strategy of the ETM has allowed us to identify two main macro-categories as areas of application of AI, as diagnosis and treatment, by highlighting the focus on the patient. The SVM classification model has also proved to be effective in testing the classification of documents based on ETM categories. However, without a semantic-oriented approach to the text, we would not have been able to improve the classification model, as we would have been unable to capture the semantic nuances that mixed diagnosis and care within

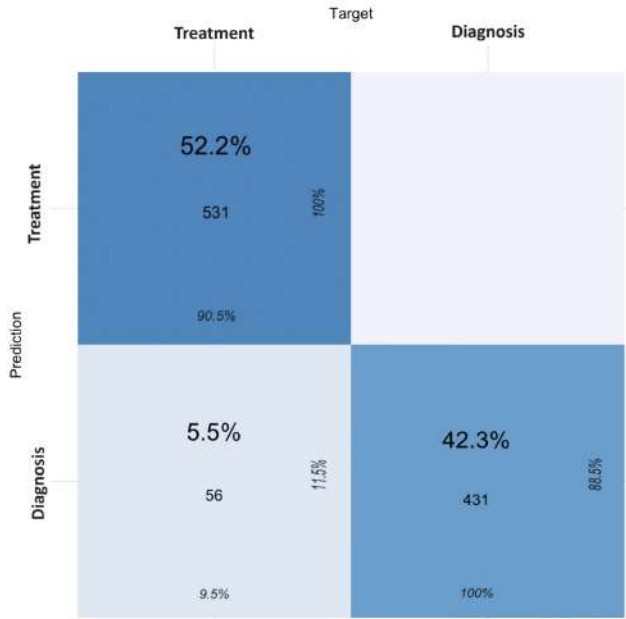


Fig. 3. Confusion Matrix (SVM classification) on sub-group of texts

the same scientific contributions. This preliminary step has, therefore, allowed us to improve the effectiveness of the SVM model with optimal performance. As a further step, we aim to obtain a more precise estimation of the third category, the one in which we have documents belonging to both diagnosis and treatment, or to propose effectively a fuzzy classifier.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003)

2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 93–1022 (2003)

3. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995)

4. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: Topic modeling embedding spaces. *Trans. Assoc. Comput. Linguist.* **8**, 439–453 (2020)

5. Kaur, A., Garg, R., Gupta, P.: Challenges facing AI and big data for resource-poor healthcare system. In: 2nd International Conference on Electronics and Sustainable Communication Systems, pp. 1426–1433 (2021)

6. Strachna, O., Asan, O.: Reengineering clinical decision support systems for artificial intelligence. In: IEEE International Conference on Healthcare Informatics, pp. 1–3 (2020)

7. Thorsten, J.: Text categorization with support vector machines: learning with many relevant features. In: Machine Learning: ECML-98. LNCS. Springer (1998)
8. Yu, H., Zhou, Z.: Optimization of IoT-based artificial intelligence assisted telemedicine health analysis system. *IEEE Access* **9**, 85034–85048 (2021)



Hidden Markov Graphical Models with Generalized Hyperbolic Distributions: A Financial Analysis on Commodities and Green Energy Indexes

Beatrice Foroni¹(✉), Luca Merlo², and Lea Petrella³

¹ Department of Management and Economics, University of Pisa, Pisa, Italy
`beatrice.foroni@ec.unipi.it`

² Department of Human Sciences, European University of Rome, Rome, Italy
`luca.merlo@uniroma1.it`

³ MEMOTEF Department, Sapienza University of Rome, Rome, Italy
`lea.petrella@uniroma1.it`

Abstract. This paper introduces a novel hidden Markov graphical model to explore temporal interconnections among diverse financial markets. Exploiting the properties of the generalized hyperbolic family of distributions with time-varying parameters governed by a latent Markov chain, we can establish conditional correlation structures among variables. Our approach employs a penalized EM algorithm with an L_1 penalty of the off-diagonal elements of the precision matrix. We apply the model to daily returns of metal, energy commodities and green energy ETFs from 2017 to 2024, showcasing its practical utility in unveiling time-varying network dynamics.

Keywords: EM algorithm · Generalized Hyperbolic family · Financial networks · Hidden Markov models

1 Introduction

In recent years, the urge to identify how the impact of financial stress events can spread across the whole financial global system has made network science a useful tool for describing the propagation of systemic risk. Within this literature, Gaussian Graphical Models (GGM) have received an enormous attention because they provide a simple method to model the pair-wise conditional dependence structure of a collection of stochastic variables. In a high-dimensional framework, when a large set of random variables is considered, we are interested in identifying only a smaller subset of variables that exhibits the most relevant and strongest dependencies. Among the several methods proposed in literature, there is the *Graphical LASSO* (*glasso*) algorithm of [5], which maximizes the

likelihood of the model penalized by the L_1 -norm of the elements of the precision matrix. The literature regarding non-Gaussian graphical models is fairly limited. Furthermore, financial time series exhibit well-known stylized facts such as fat tails, leptokurtosis and departures from normality, necessitating thorough consideration in the analysis. Particularly, clean energy stocks have emerged as a new asset class for market participants, especially environmentally responsible investors. More recently, the rise of green bond financing for climate and environmental initiatives has sparked a peak in the sector, and the high volatility of renewable energy assets has increased interest among risk managers and regulators in understanding how hidden variables, such as market conditions, impact the temporal evolution and clustering of return volatility. In this context, Hidden Markov Models (HMMs) have been successfully employed in the analysis of financial time series data, with applications to asset allocation and stock returns, [4]. Regarding undirected graphs, the estimation of graphical models within HMMs has been explored by [6]. They employed multivariate Gaussian emission distributions with sparse precision matrices, representing state-specific conditional independence graphs. A more recent contribution by [1] introduced a Markov switching graphical seemingly unrelated regression model. This model delves into time-varying systemic risk by leveraging a variety of multi-factor asset pricing models. In this work, in order to model time-varying conditional correlation structures in multivariate time series data, we consider multivariate symmetric generalized hyperbolic (GH) distributions with time-dependent parameters evolving according to a discrete, homogeneous latent Markov chain. In financial literature, the GH distribution has attracted considerable interest not only for its ability to capture essential features of return distributions but also for its substantial flexibility in modeling financial data. This flexibility encompasses various specific cases such as multivariate Normal, t-distribution, Laplace distribution, and several others. Following [3], we demonstrate that, conditionally on each latent state, the inverse of the state-specific scale matrix of the multivariate GH completely characterizes the conditional correlation structure among the random variables. To induce sparsity in the inverse of the scale matrices, i.e., the precision matrices, and identify whether two nodes are connected by an edge, we exploit the Gaussian location-scale mixture representation of the GH family to build a suitable penalized expectation-conditional maximization either (ECME) algorithm. This enables us to include in our estimation procedure the *glasso* approach of [5] accounting for an L_1 penalty on the off-diagonal elements of the precision matrices in the M-step of the algorithm. We call this method *hidden Markov generalized hyperbolic graphical model* (HMGHGM). The relevance of our approach is shown empirically on a collection of metal, energy commodities and green energy ETFs from 2017 to 2024, and the modeling approach we propose is able to identify the connectedness when different levels of distress are considered, and can describe the topological structure of the underlying graph without relying on the hypothesis of normality for the time series returns.

2 Model Specification

Formally, let $\mathbf{Y}_t = [Y_t^{(1)}, \dots, Y_t^{(d)}]'$ denote a continuous d -dimensional random vector for $t = 1, \dots, T$. The joint probability density function of $\mathbf{Y}_t$ following the (symmetric) GH distribution can be written as

$$f_{\mathbf{Y}_t}(\mathbf{y}_t; \cdot) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2} K_\lambda(\sqrt{\psi\chi})} \left[\frac{\chi + \delta(\mathbf{y}_t; \mu, \Sigma)}{\psi} \right]^{\frac{\lambda - d/2}{2}} K_{\lambda - \frac{d}{2}} \left(\sqrt{[\chi + \delta(\mathbf{y}_t; \mu, \Sigma)] \psi} \right) \quad (1)$$

where $\mu \in \mathbb{R}^d$ is the location parameter, Σ is a $d \times d$ positive definite and symmetric scale matrix, such that $|\Sigma| = 1$ for identifiability purposes, $\lambda \in \mathbb{R}$ is the index parameter, $\chi > 0$ and $\psi > 0$ are concentration parameters, $\delta(\mathbf{y}_t; \mu, \Sigma) = (\mathbf{y}_t - \mu)' \Sigma^{-1} (\mathbf{y}_t - \mu)$ is the squared Mahalanobis distance between $\mathbf{y}_t$ and μ with scale matrix Σ and finally $K_{\lambda - \frac{d}{2}}(\cdot)$ denotes the modified Bessel function of the third kind of order $\lambda - \frac{d}{2}$. We adopt the compact notation $\mathbf{Y}_t \sim \mathcal{GH}_d(\mu, \Sigma, \lambda, \chi, \psi)$. One of the key benefits of the GH distribution is that, using (1), $\mathbf{Y}_t$ admits the following location-scale mixture representation:

$$\mathbf{Y}_t = \mu + \sqrt{W_t} \Sigma^{1/2} \mathbf{Z}_t \quad (2)$$

where $\mathbf{Z}_t \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ denotes a d -variate standard Normal distribution whose covariance matrix is the identity matrix and W_t has a generalized inverse Gaussian (GIG) distribution, $W_t \sim \mathcal{GIG}(\lambda, \chi, \psi)$, with $\mathbf{Z}_t$ being independent of W_t . The GH family encompasses several well-known, applied models for financial data by varying appropriately the values of the parameters of the GIG distribution, (λ, χ, ψ) , including the multivariate Laplace, t and Normal distribution.

Now let's describe the proposed hidden Markov graphical model with state-specific multivariate GH emission densities. Let $\{S_t\}_{t=1}^T$ be a latent, homogeneous, first-order Markov chain defined on the discrete state space $\{1, \dots, K\}$. Let $\pi_k = Pr(S_1 = k)$ be the initial probability of state k , $k = 1, \dots, K$, and $\pi_{k|j} = Pr(S_{t+1} = k | S_t = j)$, with $\sum_{k=1}^K \pi_{k|j} = 1$ and $\pi_{k|j} \geq 0$, denote the transition probability between states j and k , that is, the probability to visit state k at time $t+1$ from state j at time t , $j, k = 1, \dots, K$ and $t = 1, \dots, T$. We collect the initial and transition probabilities in the K -dimensional vector π and in the $K \times K$ matrix Π , respectively.

In our model we assume that the conditional distribution of $\mathbf{Y}_t$ given the state occupied by the latent process at time t corresponds to the GH distribution in (1), whose parameters depend on the value of the Markov chain S_t , namely $\mathbf{Y}_t | S_t = k \sim \mathcal{GH}_d(\mu_k, \Sigma_k, \lambda_k, \chi_k, \psi_k)$. We define with $\Theta_k = \Sigma_k^{-1}$ for $k = 1, \dots, K$, the precision matrices used to build state-specific undirected graphs that conveys the conditional intercorrelation structure among the elements of $\mathbf{Y}_t$ given the latent process S_t . More precisely, supposing $S_t = k$, let $G_k = (V, E_k)$ be an undirected state-dependent graph where $V = \{1, \dots, d\}$ denotes the set of state-invariant nodes, such that each component of the random variable $\mathbf{Y}_t$

corresponds to a node in V , and $E_k \subseteq V \times V$ represents the set of undirected edges in the k -th state. In order to study the conditional correlation structure of $\mathbf{Y}_t$ within the k -th state through the graph G_k , we establish a useful result that allows us to make inference on the edge set E_k based on the precision matrix Θ_k . Following [3] and exploiting the mixture representation of the symmetric GH in (2), we can characterize the conditional correlation structure as shown in the following proposition.

Proposition 1. *Consider $S_t = k$ and let $\mathbf{Y}_t \mid S_t = k \sim \mathcal{GH}_d(\mu_k, \Sigma_k, \lambda_k, \chi_k, \psi_k)$. If two nodes j and h , with $j, h \in V$ and $j \neq h$, of the graph are separated by a set of nodes $C \in V$, then $Y_t^{(j)}$ and $Y_t^{(h)}$ are conditionally uncorrelated given $\mathbf{Y}_t^{(C)}$ and $S_t = k$.*

The proof of Proposition 1 follows directly from the mixture representation of the $\mathcal{GH}$ in (2) and the closure property of the Normal distribution under conditioning of its components. Most importantly, from Proposition 1 we can exploit the properties of GGMs to characterize conditional correlation relationships between the elements of $\mathbf{Y}_t$ in each state $k = 1, \dots, K$, estimating the precision matrices Θ_k 's. To estimate and make inference on the model parameters we develop a suitable penalized expectation-conditional maximization either (ECME) algorithm, which exploits the mixture representation of the $\mathcal{GH}$ distribution, treating W as missing data. In order to identify only a smaller subset of variables that exhibit the most relevant and strongest dependencies, an L_1 -norm penalty of the off-diagonal elements of Θ_k is added to the likelihood of the model.

The penalized version of the conditional Normal log-likelihood is reported in the following definition.

Definition 1. *The conditional Normal log-likelihood function is proportional to:*

$$\ell_{pen}(\mu_k, \Theta_k \mid \mathbf{Y}_t, W_{tk}, S_t = k) \propto \frac{1}{2} \sum_{t=1}^T \gamma_t(k) \log |\Theta_k| - \frac{1}{2} \text{tr}\{\tilde{\mathbf{S}}_k \Theta_k\} - \rho \sqrt{\nu_k} \|\Theta_k\|_1 \quad (3)$$

where $\|\Theta_k\|_1$ is the sum of the absolute values of the off-diagonal entries of the matrix Θ_k and $\tilde{\mathbf{S}}_k$ is the weighted empirical covariance matrix defined as

$$\tilde{\mathbf{S}}_k = \sum_{t=1}^T \frac{\gamma_t(k)}{W_{tk}} (\mathbf{Y}_t - \mu_k)(\mathbf{Y}_t - \mu_k)'. \quad (4)$$

Additionally, $\nu_k > 0$ is a state-dependent weight to allow for a different penalty in each latent state such that $\sum_{k=1}^K \nu_k = 1$ and $\rho \geq 0$ is the tuning parameter.

As one can see, the penalized complete likelihood in (3) is exactly the objective function maximized in a GGM where the empirical covariance matrix of the data is substituted by the matrix $\tilde{\mathbf{S}}_k$ defined in (4), for $k = 1, \dots, K$. This suggests that we can exploit the simple and fast *lasso* algorithm of [5] to update the estimate of Θ_k .

3 Real Data Results and Conclusions

The empirical analysis is performed on the log-returns of 15 financial assets comprising metal and energy commodities and green energy ETFs from November 3, 2017 to February 12, 2024, for a total of $T = 1554$ observations after removing missing data. In order to select the optimal ρ in (3) and number of latent states K simultaneously, we fit the proposed HMGHGM for a sequence of 300 values of ρ for $K = \{1, \dots, 4\}$ and select the pair (ρ, K) using the Mixture Minimum Description Length (MMDL), which chooses $K = 3$.

Table 1. Commodities and ETFs considered in the analysis. Numbers identify vertex labels in Fig. 2.

Metal commodities	Energy commodities	Green Energy ETF
1. Gold	6. Crude oil	11. Invesco Global Clean Energy (PBD)
2. Silver	7. Heating oil	12. Invesco WilderHill Clean Energy (PBW)
3. Platinum	8. Natural gas	13. Amundi Index Solutions (CWE.MI)
4. Copper	9. RBOB Gasoline	14. First Trust NASDAQ Clean Edge Green Energy Index Fund
5. Palladium	10. Brent Crude Oil	15. S&P Green Bond

Figure 1 reports the predicted posterior probabilities of hidden states over time with $K = 3$ states. Specifically, State 3, spanning from late 2017 to the financial market collapse in March 2020 due to the COVID-19 pandemic, corresponds to a phase of low prices. State 2 reflects the growth in green energy ETFs prices throughout the 2020 and the beginning of 2021, highlighting the market's expectations of the potential victory of Joe Biden's campaign strongly focused on environmental policies. Finally, State 1 is characterized by the price growth of energy commodities, Natural Gas in particular, during 2022, as direct consequence of the Russian-Ukrainian war [2].

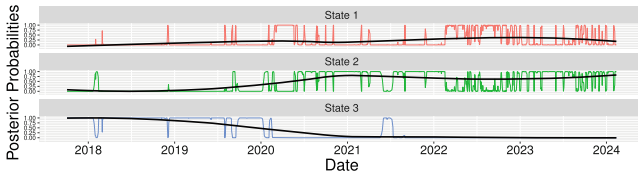


Fig. 1. Predicted posterior probabilities of hidden states over time with $K = 3$ states. The over-imposed black lines denote the trend estimated by a local linear regression.

Examining the estimated state-dependent graphs (see Fig. 2) we see distinct network connectivity and strong clustering patterns among commodities and indexes of the same sectors. The strong positive connections among WTI Crude Oil and Brent Crude Oil, as well as Gold and Silver or Invesco WilderHill and NASDAQ Clean Energy Index Fund define strong correlations in every market

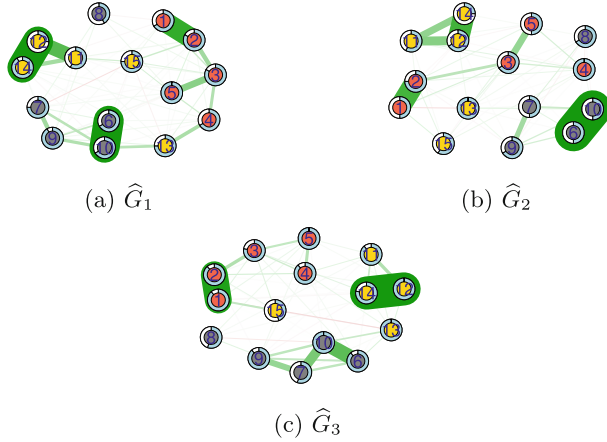


Fig. 2. Estimated graphs for $k = 1$ (a), $k = 2$ (b) and $k = 3$ (c). Red, grey and yellow nodes represent respectively metals, energy commodities and green energy ETFs while the vertex labels are illustrated in Table 1. Green edges in the networks depict positive associations, while red edges represent negative associations. Pies in light blue indicate degree centrality. The greater the pie, the higher is the degree centrality. (Color figure online)

condition. Natural Gas appears to be, especially in the State 1, the least connected node in the network. This behavior allows us to highlight once more the anomaly of its price increase throughout 2022. Moreover, the centrality of the Amundi Index Solutions ETF highlights how this index was impacted by the “green boom” of 2020. In conclusion, this model not only strengthens but also expands the existing literature in this field by introducing a technique to adapt a *Hidden Markov glasso* algorithm to non-Gaussian distributions.

References

1. Bianchi, D., Billio, M., Casarin, R., Guidolin, M.: Modeling systemic risk with Markov switching graphical SUR models. *J. Econ.* **210**(1), 58–74 (2019)
2. D’Ecclesia, R.L., Morelli, G., Stefanelli, K.: Energy ETF performance: the role of fossil fuels. *Energy Econ.*, 107332 (2024)
3. Finegold, M., Drton, M.: Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.*, 1057–1080 (2011)
4. Foroni, B., Merlo, L., Petrella, L.: Expectile hidden Markov regression models for analyzing cryptocurrency returns. *Stat. Comput.* **34**(2), 66 (2024)
5. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9**(3), 432–441 (2008)
6. Städler, N., Mukherjee, S.: Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *Ann. Appl. Stat.*, 2157–2179 (2013)



L_0 -Penalized Membership in Sparse Fuzzy Clustering

Maria Brigida Ferraro, Marco Forti^(✉), and Paolo Giordani

Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
{mariabrigida.ferraro,marco.forti,paolo.giordani}@uniroma1.it

Abstract. The literature encompasses various approaches to address sparsity in the context of cluster analysis, often by adding regularization terms to weight the role of variables in clustering processes. In the work presented here, instead, an L_0 -regularization term is employed on fuzzy membership degrees to enhance the standard Fuzzy K -Means algorithm. The new algorithm helps to assign units very close to the corresponding prototype with a membership degree equal to 1 without necessarily compromising the potential ambiguity in the membership of some units.

Keywords: Clustering · Fuzzy K -Means · Sparsity · L_0 -regularization

1 Introduction

The standard clustering algorithms aim to partition a set of units into clusters, where each unit can be assigned to only one cluster (see, for example, [1]). In the field of cluster analysis, we can distinguish between algorithms that strictly assign units to clusters (hard clustering, e.g., K -Means (KM) [2]) and algorithms that recognize this membership with a certain degree ranging between 0 and 1 (fuzzy clustering, e.g., Fuzzy K -Means (FKM) [3]).

Our work contributes to this literature by introducing an L_0 -regularization term in the objective function to be minimized. Specifically, the regularization concerns membership degrees, allowing sparsity in the FKM solution, i.e., membership degrees equal to 1 when the membership of units to clusters is evident, while not excluding the possibility of fuzzy membership when needed.

2 Fuzzy k -Means Clustering Algorithm

The Fuzzy K -Means (FKM) clustering algorithm [3] aims to find the best fuzzy partition of N units into K clusters based on S variables. Defining $\mathbf{X}$ as the data matrix of order $(N \times S)$, where each row $\mathbf{x}_i = [x_{i1}, \dots, x_{iS}]$ refers to the i -th unit ($i = 1, \dots, N$), FKM involves solving the following minimization problem:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{P}} J_{\text{FKM}} &= \sum_{i=1}^N \sum_{g=1}^K u_{ig}^m d^2(\mathbf{x}_i, \mathbf{p}_g), \\ \text{s.t. } u_{ig} &\in [0, 1], i = 1, \dots, N, g = 1, \dots, K, \\ \sum_{g=1}^K u_{ig} &= 1, g = 1, \dots, K. \end{aligned} \quad (1)$$

where $d(\cdot)$ denotes the Euclidean distance. The result of this approach is the definition of a matrix $\mathbf{P}$ of order $(K \times S)$ containing prototypes, where each row $\mathbf{p}_g = [h_{g1}, \dots, h_{gS}]$ refers to the prototype of the g -th cluster ($g = 1, \dots, K$), and a matrix $\mathbf{U}$ - $(N \times K)$ - whose generic element $u_{ig} \in [0, 1]$ represents the membership degree of unit i to cluster g .

The use of regularization terms in clustering, including possibilistic or model-based approaches, has been explored by several authors in recent years (see e.g. [4–18]). Proposals often involve penalty terms for variable selection. In our case, namely in Fuzzy K -Means with L_0 -penalty (FKM- L_0), regularization on the membership degree matrix leads to a reformulation of the problem presented in (1) as follows:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{P}} J_{\text{FKM-}L_0} &= \sum_{i=1}^N \sum_{g=1}^K u_{ig}^m d^2(\mathbf{x}_i, \mathbf{p}_g) + \lambda \sum_{i=1}^N \|\mathbf{u}_i\|_0, \\ \text{s.t. } u_{ig} &\in [0, 1], i = 1, \dots, N, g = 1, \dots, K, \\ \sum_{g=1}^K u_{ig} &= 1, g = 1, \dots, K. \end{aligned} \quad (2)$$

$\|\cdot\|_0$ regularization works by encouraging the coefficients of the model to take on either 0 or 1 values. The L_0 -norm of a vector $\mathbf{y}$ of length n is defined as the number of non-zero elements in the vector which can be expressed as:

$$\|\mathbf{y}\|_0 = \sum_{i=1}^n \mathbb{I}_{\{y_i \neq 0\}}$$

where $\mathbb{I}_{\{y_j \neq 0\}}$ is the indicator function that returns 1 if y_i is non-zero and 0 otherwise. In other words, the L_0 norm counts the number of non-zero elements in the vector $\mathbf{y}$ by summing the indicator functions, which are equal to 1 for each non-zero element and 0 otherwise.

Operatively speaking, the algorithm aims to replace, for $i = 1, \dots, N$, the element characterized by the lowest membership with zero and then evaluate the goodness of the replacement in terms of the loss function. In this context it should be emphasized that the algorithm tends toward classical FKM for $\lambda = 0$, which represents the zero penalty situation. It is hence straightforward to see how as the values of λ increase, a defuzzification of $\mathbf{U}$ is achieved. In other words, the membership degree matrix $\mathbf{U}$ tends to become sparse because, when units clearly belong to the clusters, the FKM- L_0 algorithm assigns membership degrees of 1 to such units in the involved clusters and, obviously, 0's with respect to the remaining clusters. In line with the fuzzy approach, units with unclear memberships are assigned to clusters with degrees in the range $[0, 1]$.

3 Synthetic Example

To simulate the behavior of the algorithm, we generated synthetic data representing four well-separated clusters, each containing five clearly assigned units and four units with less distinct memberships. A visual representation of the generated dataset can be observed in Fig. 1.

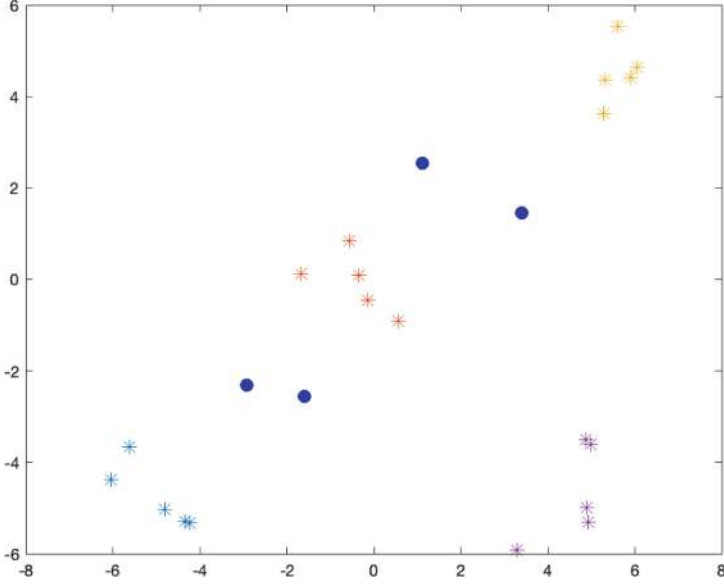


Fig. 1. Toy example

In Table 1, the membership degrees obtained from FKM algorithm (on the left) and those obtained from a penalized FKM- L_0 algorithm (on the right) are reported. For this experiment we adopted a default value of 2 for the fuzziness parameter (m), while the penalty value (λ) was set equal to the ratio of the loss function obtained by the classic FKM method and the number of observations N which, in this particular example, resulted to be equal to 1.5489. As observed, the first algorithm suggests a purely fuzzy partition even when units clearly belong to a cluster: in fact, a certain degree of membership is assigned for each cluster. On the other hand, the proposed algorithm manages to unambiguously assign various units in line with the expectations, without disregarding the possibility of fuzziness.

Table 1. Membership degrees matrices obtained from FKM (on the left) and FKM- L_0 (on the right)

FKM				FKM- L_0			
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.020	0.005	0.965	0.011	0	0	1	0
0.034	0.007	0.945	0.013	0	0	1	0
0.029	0.007	0.951	0.013	0	0	1	0
0.005	0.001	0.992	0.002	0	0	1	0
0.017	0.004	0.971	0.009	0	0	1	0
0.953	0.016	0.017	0.014	1	0	0	0
0.891	0.027	0.034	0.048	1	0	0	0
0.984	0.004	0.006	0.006	1	0	0	0
0.901	0.023	0.050	0.026	1	0	0	0
0.999	0.000	0.001	0.000	1	0	0	0
0.001	0.999	0.000	0.000	0	1	0	0
0.006	0.987	0.002	0.005	0	1	0	0
0.003	0.994	0.001	0.002	0	1	0	0
0.012	0.976	0.003	0.008	0	1	0	0
0.020	0.960	0.006	0.013	0	1	0	0
0.063	0.027	0.045	0.864	0	0	0	1
0.011	0.006	0.006	0.977	0	0	0	1
0.030	0.019	0.012	0.939	0	0	0	1
0.025	0.016	0.011	0.948	0	0	0	1
0.005	0.003	0.003	0.990	0	0	0	1
0.552	0.044	0.302	0.102	0.639	0	0.361	0
0.377	0.038	0.514	0.071	0.413	0	0.587	0
0.350	0.444	0.055	0.151	0.434	0.566	0	0
0.612	0.239	0.063	0.086	0.716	0.284	0	0

4 Conclusion

In this paper, we presented a novel approach which implements the Fuzzy K -Means clustering algorithm with L_0 -regularization in the membership degree matrix. Thanks to the L_0 -regularization term, the algorithm tends to produce hard assignment (membership degrees equal to 1) for the units *clearly* belonging to the clusters without eventually losing relevant information in case of units with fuzzy assignments.

References

1. Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)
2. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 281–298 (1967)
3. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Springer, New York, NY (1981)
4. Inokuchi, R., Miyamoto, S.: Sparse possibilistic clustering with L1 regularization. In: Proceedings IEEE International Conference on Granular Computing, pp. 442–445 (2007)
5. Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8**, 1145–1164 (2007)
6. Namkoong, Y., Heo, G., Woo, Y.: An extension of possibilistic fuzzy c -means with regularization. In: Proceedings IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6 (2010)
7. Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. *J. Am. Stat. Assoc.* **10**, 713–726 (2010)
8. Qiu, X., Qiu, Y., Feng, G., Li, P.: A sparse fuzzy c -means algorithm base on sparse clustering framework. *Neurocomputing* **157**, 290–295 (2015)
9. Arias-Castro, E., Pu, X.: A simple approach to sparse clustering. *Comput. Stat. Data Anal.* **105**, 217–228 (2017)
10. Guo, L., Chen, L., Lu, X., Chen, C.L.P.: Membership affinity lasso for fuzzy clustering. *IEEE Trans. Fuzzy Syst.* **28**, 294–307 (2020)
11. Yang, M.-S., Ali, W.: Fuzzy Gaussian Lasso clustering with application to cancer data. *Math. Biosci. Eng.* **17**, 250–265 (2020)
12. Yang, M.-S., Benjamin, J.B.M.: Feature-weighted possibilistic c -means clustering with a feature-reduction framework. *IEEE Trans. Fuzzy Syst.* **29**, 1093–1106 (2021)
13. Benjamin, J.B.M., Yang, M.-S.: Weighted multiview possibilistic C -means clustering with L2 regularization. *IEEE Trans. Fuzzy Syst.* **30**, 1357–1370 (2022)
14. Wang, J., Ma, Z., Nie, F., Li, X.: Entropy regularization for unsupervised clustering with adaptive neighbors. *Pattern Recognit.* **125**, 108517 (2022)
15. Zhao, X., Nie, F., Wang, R., Li, X.: Improving projected fuzzy K -means clustering via robust learning. *Neurocomputing* **491**, 34–43 (2022)
16. Chen, J., Zhu, J., Jiang, H., Yang, H., Nie, F.: Sparsity fuzzy C -means clustering with principal component analysis embedding. *IEEE Trans. Fuzzy Syst.* **31**, 2099–2111 (2023)
17. Yang, M.-S., Benjamin, J.B.M.: Sparse possibilistic c -means clustering with Lasso. *Pattern Recognit.* **138**, 109348 (2023)
18. Wang, J., Zhang, X., Nie, F., Li, X.: Enhanced robust fuzzy K -means clustering joint l_0 -norm constraint. *Neurocomputing* **561**, 126842 (2023)



Secondary Migration Intentions of Refugees Living in Germany

Cecilia Fortunato¹(✉), Hans Dietrich², and Elena Ambrosetti¹

¹ Sapienza University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy
cecilia.fortunato@uniroma1.it

² Institute for Employment Research (IAB), Regensburger Street 104,
90478 Nürnberg, Germany

Abstract. A crucial topic in Migration Studies is the exploration of refugees and other migrants' mobility aspirations and behaviors over the life course. Beyond micro-trajectories and individual characteristics, a central role is played by the uncertainty connected to political, social and cultural contexts and to structural conditions both in native and host country. Using original data from the Refugee Sample of the German Socio-Economic Panel and performing logistic regression analyses, we investigate the intention to stay in Germany permanently, the intention to leave and the uncertainty about future mobility choices of 7,134 adult refugees arrived from Syria, Iraq, Afghanistan, Eritrea and Iran after 2011. For what concerns conditions in destination, in line with previous studies, we found that factors shaping onward migration or return intention are mainly related to labor market integration and mismatch with qualification and linguistic integration, along with family members' place of residence. Nevertheless, the main contribution of this paper can be claimed for the relevant role played by soft indicators such as well-being measures, worries and concerns about the future, triggering uncertainty in refugees' mobility choices linked to conditions in origin country, legal and policy evolution and public discourses about ethnic minorities in destination country.

Keywords: Refugees return · Onward migration · Return migration

1 Context and Theoretical Background

Germany has a long history as main destination for international migration flows and assumed a leading role in European asylum reception system in 2015 and 2016, when historical peaks of more than one million refugees applied for the international protection [2, 7]. Germany has been the main European destination for refugees from Syria since the outbreak of the civil war in March 2011 (765,000 new protection claims in 2015–2016) and registered the record number of first-time asylum applications of migrants from Afghanistan (154,000), Iraq (122,000), Eritrea (33,400) and Iran (40,160) [7]. A crucial yet understudied area in the field of migration research is the complex decision-making process that drives refugees and other migrants to settle permanently in the host country, or move onward to a third destination, or voluntarily and permanently return

backward to native country [3, 5, 6]. Several challenges are faced by researchers in the estimation of outward migration and return intentions and realizations, mainly related to the lack of consistent and updated register and survey data [5, 6]. In the context of large-scale displacements of people fleeing war-torn zones, understanding refugees' decision-making process implies considering a complex combination of micro level factors (individual characteristics, background and integration) and macro level factors (economic and institutional contexts) [10], often entrenched within great uncertainty [1, 9, 11]. All migrants have a certain degree of agency and freedom over their mobility choices [4]. Nevertheless, displaced refugees may experience greater uncertainty about contexts' evolution, lack of concrete alternatives, limited capability to aspire the return to home [4, 9]. As refugees seek international protection in other countries in order to ensure their safety, they are typically unable or unwilling to return to their home country unless the conditions precipitating their flight have ended and the situation at home has improved significantly [1, 9, 11]. Outlining the different paths refugees may choose following displacement and understanding refugees' aspirations for stay, onward migration or return is essential for designing effective policies that facilitate integration in the host country or assisted voluntary repatriation and successful reintegration programs in home country.

2 Aim of the Study

The present study aims at contributing to the advancement of the literature on refugees' secondary migration intentions, providing new evidence from the Survey of Refugees carried out in Germany from 2016 on. We focus on the stay/leave intention and uncertainty expressed by respondents in their first interview. We attempt to understand the complex interplay of different factors connected to individual traits and migration experiences, socio-economic characteristics and subjective well-being dimensions linked to worries and hopes about the future.

The following research questions guided our study:

- i) What are the characteristics of refugees arrived in Germany after 2011 and expressing the intention to leave or to stay in the German Survey of Refugees?
- ii) What factors influence the intention to leave and trigger uncertainty about the opportunity to settle permanently in Germany?

3 Data and Method

3.1 Sample from the Survey of Refugees 2016–2020

Data stem from the German nationwide household and individual IAB-BAMF-SOEP Survey of Refugees, carried out from 2016 on by a joint research group¹ [8]. The first wave was conducted in 2016 with asylum seekers and refugees migrated to Germany

¹ i) Institute for Employment Research (IAB); ii) Integration and Asylum Research Centre at the Federal Office for Migration and Refugees (BAMF-FZ); iii) German Socio-Economic Panel (SOEP). <https://doi.org/10.5684/soep.core.v37eu>.

between 2013 and 2016, integrated by further sample's refreshments and enlargements of refugees arrived between 2017 and 2019 [8]. All combined, the samples are representative of the population of refugees and asylum seekers arrived in Germany between 2013 and 2019 and registered in the Central Register of Foreigners by the year 2020. We select the five countries of origin with higher numbers of respondents in the survey, corresponding to the national states from which major flows of refugees have sought asylum in Germany during the period considered. We obtain a final sample of 20,609 observations for 4,712 households and 7,453 individuals. The sample includes individuals migrated from Syria ($n = 4,518$, 60.6%), Iraq ($n = 1,202$, 16.1%), Afghanistan ($n = 1,028$, 13.8%), Eritrea ($n = 438$, 5.9%) and Iran ($n = 267$, 3.6%).

3.2 Variables of Interest

Dependent variable: In the first wave of the survey (2016–2017), refugees' intentions are explored through the question “Would you like to stay in Germany permanently?”. The possible answers were “Yes”, “No”, “Not sure”. For ease of interpretation and for a clearer connection to our research questions, we flip the issue into the intention to leave Germany, recoding the variable as “0. No, want to stay permanently in Germany”, “1. Yes, I want to leave” “99. Not sure”.

Predictors: i) *Socio-demographic:* gender, age, country of birth; ii) *Migration background:* migration drivers; protection and residence status; duration of stay; iii) *Economic and labor market integration:* education; employment; linguistic integration; iv) *Networks and family:* household members' information and residence; v) *Subjective Wellbeing:* health status; life satisfaction; disadvantages, discrimination, worries and hopes about futures.

3.3 Method

Given a set of multi-dimensional predictors, we develop a multinomial logistic regression analysis, controlling for socio-demographic variables (gender and age): we simultaneously analyze the stated intention to leave (1) in opposition to the intention to stay and the “not sure” status (2) against the intention to stay (0).

4 Results

4.1 Descriptive Statistics

We analyze the main characteristics of the subsample providing an answer to the question about stay/leave intention (RQ1) in the first interview ($n = 7,134$)². Answers are unequally distributed across categories, as 90.9% of respondents expresses the intention to stay permanently in Germany ($n = 6,481$), 4.6% expresses the intention to leave ($n = 331$) and 4.5% expresses uncertainty about future mobility choices ($n = 322$). Considering gender and age of respondents, stay intention is more frequent among females

² Information about the total sample are available upon request to the authors.

(91.3%, against 90.5% of males), and respondents in older ages (93.7% for age group 54–65, averagely 90.0% for age groups 18–34 and 35–54). Considering all national groups, the great majority expresses the intention to stay, ranging from 86.8% for Syrian up to 97.8% for Eritrean. Nevertheless, Syrian refugees are more likely to express leaving intention (6.7%) and be affected by uncertainty (6.6%) compared to other groups (ranging from 0.7% up to 2.3%). The length of stay is positively associated with the intention to stay, as 97.4% of individuals who arrived more than five years earlier expresses the will to stay, against the 89.7% of individuals who arrived one year earlier or less. Looking at the protection status, a lower percentage of recognized refugees express the intention to stay (89.5%), compared to respondents still waiting for the asylum application outcome (95.4%) and respondents for whom the international protection has been denied (94.8%). Finally, analyzing socio-economic aspect, uncertainty about future or the intention to leave is expressed by higher percentage of individuals holding a university degree (17.1%), as well as individuals in education programs (12.1%) or seeking a job (10.6%), compared to individuals with no professional education (7.4%), or having permanent (8.3%) or temporary job contracts (8.2%).³

4.2 Regression Analysis

We then analyze simultaneously factors influencing the intention to leave and triggering uncertainty about the opportunity to settle permanently in Germany (RQ2). Relative risk ratios, holding all other variables constant, show the ratio of the probability of choosing a certain category of the dependent variable, over the probability of choosing the base category, for a one-unit increase in the predictor variable (Table 1). Looking at the country of origin, all national groups present a lower intention to leave or lower uncertainty compared to Syrians. The difference becomes particularly significant with respect to refugees from Afghanistan. We confirm a positive effect of the length of stay on permanent settlement, as living in Germany for more than five years negatively affects leaving intention and uncertainty, compared to refugees arrived in the previous four years. Looking at household's characteristics, a slight negative effect on onward migration intention and uncertainty can be seen for respondents with partner living in Germany, whilst contrary to what we would expect, living in a private apartment has a positive effect, compared to respondents living in shared accommodations for refugees. In compliance with previous studies, refugees holding a university degree are more likely to express the intention to leave compared to refugees with lower educational attainment and are far more likely to be uncertain about future, as well as refugees self-rating a very good level in speaking German. These results may reflect labor and housing market discriminating practices discouraging refugees stay intentions. Not surprisingly, life satisfaction score is negatively associated to the intention to leave, whilst feeling disadvantaged due to ethnic origin and worries about hostility against foreigners positively affect the intention to leave. On origin's side, the intention to leave is strongly connected to the hope for the end of armed conflicts as a necessary condition to be able to return, considered more relevant than changes in political, institutional and economic contexts. These worries,

³ Other descriptive results are available upon request to the authors.

Table 1. Relative Risk Ratios from Multinomial Logistic Regression model: intention to leave (1) and not sure (2) of refugees living in Germany, baseline is the intention to stay (0)

n = 7,011	Intention to leave (1)		Not sure (2)	
Predictors	RRR (95% CI)	St. Err	RRR (95% CI)	St. Err
Country origin (ref. 1. Syria)				
2. Iraq	0.31 ***	0.09	0.16 ***	0.05
3. Afghanistan	0.43 ***	0.11	0.26 ***	0.08
4. Eritrea	0.21 ***	0.13	0.31 ***	0.13
5. Iran	0.60	0.27	0.37 **	0.16
Duration of stay (ref. 1. One year or less)				
2. More than one, less than five years	1.17	0.15	0.67 ***	0.09
3. Five years or more	0.25 ***	0.12	0.23 ***	0.10
Partner residence (ref. 0. No partner)				
1. In Germany	0.90	0.13	0.48 ***	0.07
2. Abroad	0.78	0.20	0.41 ***	0.11
Type of accommodation (ref. 0. Private accommodation)				
1. Shared accommodation	0.58 ***	0.09	0.80	0.12
Educational attainment (ref. 1. No professional educational)				
2. Vocational	1.08	0.31	1.53	0.41
3. University	1.56 ***	0.22	1.76 ***	0.25
German speaking (ref. 1. Not at all)				
2. Not very well	0.99	0.22	1.48	0.40
3. Averagely	0.87	0.19	1.51	0.41
4. Well	0.87	0.22	1.46	0.43
5. Very well	1.24	0.45	2.69 ***	1.00
Life satisfaction (ref. 1. Not at all)				
2. Partly satisfied	0.60	0.19	0.70	0.30
3. Satisfied	0.49 ***	0.13	1.16	0.40
4. More than satisfied	0.28 ***	0.07	0.91	0.31
5. Very satisfied	0.24 ***	0.07	0.73	0.26
Necessary condition for return (ref. 0. no selection)				

(continued)

Table 1. (continued)

n = 7,011	Intention to leave (1)		Not sure (2)	
Predictors	RRR (95% CI)	St. Err	RRR (95% CI)	St. Err
1. End of war	3.05 ***	0.74	1.64 ***	0.21
Disadvantage due to ethnicity (ref. 0. no selection)				
1. Has felt disadvantaged	1.38 **	0.18	1.63 ***	0.21
Worried about hostility (ref. 0. no selection)				
1. Worried about hostility	1.50 ***	0.21	1.59 ***	0.23
Worried not being able to stay (ref. 0. no selection)				
1. Worried not being able to stay	0.49 ***	0.06	0.51 ***	0.07
Worried not being able to return (ref. 0. no selection)				
1. Worried not being able to return	4.79 ***	0.72	2.92 ***	0.40
Chi-square (54)	861.281			
Prob > chi2	0.000			
Pseudo r-squared	0.167			

*** p < .01. ** p < .05. * p < .1.

Source: Authors' elaborations on the IAB-BAMF SOEP Survey of Refugees 2016–2017.

together with the fear of not being able to stay in Germany or to return to native country are the most relevant components of the uncertainty about future.

5 Conclusion

Selected results reported above persuade us to reflect on the multi-dimensional nature of secondary migration intentions and the uncertainty connected to conditions both in native and host country. In the selected sample, the great majority of respondents declared the intention to stay permanently in Germany. Likely, part of this subsample is affected by what is known in literature as “desirability” response bias, distortions due to participants’ concerns about socially acceptable responses. This is particularly relevant if we consider the extreme vulnerability of refugees as the survey target group. However, it could still be possible to reveal patterns of more than 650 respondents evenly selecting the intention to leave or expressing uncertainty about future. Socio-demographic indicators such as gender and age, the protection status and objective economic measures such as current employment status, seem to have no predictive power in determining the intention to stay or to leave. In line with previous studies for what concerns conditions in destination, we found that factors shaping onward migration or return intention can mainly be related to economic unfulfilled expectations and mismatch between education,

linguistic integration and access to the labor and housing market. Nevertheless, the main contribution can be claimed for the relevant role played by soft indicators such as wellbeing measures. Worries and hopes about the future. Triggering uncertainty in refugees' mobility choices and mainly linked to legal and policy evolution as well as public discourses about ethnic minorities. These results impose a deep reflection on how to implement and harmonize these aspects in surveys' design in the context of research on secondary migration intentions of forcibly displaced migrants' crossing international borders.

References

1. Alrababah, A., Masterson, D., Casalis, M., Hangartner, D., Weinstein, J.: The dynamics of refugee return: syrian refugees and their migration intentions. *Brit. J. Polit. Sci.* **53**(4), 1108–1131 (2023)
2. BAMF: Das Bundesamt in Zahlen 2018: Asylum. Migration und Integration. Bundesamt für Migration und Flüchtlinge. Nürnberg (2019)
3. Borjas, G.J., Bratsberg, B.: Who leaves? The outmigration of the foreign-born. *Rev. Econ. Stat.* **78**(1), 165 (1996). <https://doi.org/10.2307/2109856>
4. Carling, J.: The phrase 'refugees and migrants' undermines analysis, policy and protection. *Int. Migr.* **61**(3), 399–403 (2023)
5. Constant, A., Massey, D.S.: Self-selection, earnings, and out-migration: a longitudinal study of immigrants to Germany. *J. Popul. Econ.* **16**(4), 631–653 (2003)
6. Dustmann, C., Weiss, Y.: Return migration: theory and empirical evidence from the UK. *Brit. J. Industrial Rel.* **45**(2), 236–256 (2007)
7. EUROSTAT: Asylum in the EU Member States: 1.2 million first time asylum seekers registered in 2016. Syrians. Afghans and Iraqis continued to be the top citizenships. Newsrelease No 46 (2017)
8. IAB-BAMF-SOEP Survey of Refugees 2019. data 2016–2020. <https://doi.org/10.5684/soep.iab-bamf-soep-mig.2020> (2022)
9. Müller-Funk, L.: Adapting to staying. or imagining futures elsewhere: Migration decision-making of Syrian refugees in Turkey. IMI Working Papers Series No. 155 (2019)
10. Paparusso, A., Ambrosetti, E.: To stay or to return? Return migration intentions of Moroccans in Italy. *Int. Migr.* **55**(6), 137–155 (2017)
11. Zetter, R.: Refugees and their return home: unsettling matters. *J. Refugee Stud.* **34**(1), 7–22 (2021). <https://doi.org/10.1093/jrs/feab005>



A Bayesian Binomial Regression Model for Ozone Levels in Northern Italy

Michela Frigeri, Leonardo Marchesin, Matteo Coppellotti,
and Alessandra Guglielmi^(✉)

Department of Mathematics, Politecnico di Milano, Milan, Italy
{michela.frigeri,alessandra.guglielmi}@polimi.it

Abstract. This study proposes a Bayesian binomial regression model for monthly counts of extreme ozone levels. We implement a hierarchical modeling framework, involving different hierarchies of mixed effects, to best explain pollution data provided by *ARPA Lombardia*. We consider two different daily thresholds, providing two associated datasets counting the number of days above the threshold for each month. The data are collected over multiple years through a wide network of monitoring stations. The model we propose accounts for both spatial structure and time-varying covariates in a convenient and flexible structure. We apply the same Bayesian model to the two datasets on extreme ozone pollution in Northern Italy, in particular focusing on potential differences in the posterior inference.

Keywords: Count data · ozone pollution · spatial data

1 Introduction

Ozone pollution in Northern Italy is a significant environmental concern affecting human health and ecosystems. While ozone high in the atmosphere protects humans from the sun's harmful ultraviolet rays, it is a harmful air pollutant at the ground level. Ozone (O_3) is considered a secondary pollutant, i.e., it is not emitted into the atmosphere by direct sources. Sunlight and other environmental factors provide favorable conditions for the chemical reactions that produce ozone involving primary pollutants, such as nitrogen oxides (NO_x) and volatile organic compounds (VOCs). We propose here a Bayesian modeling framework for O_3 concentrations overcoming specific thresholds. We include meteorological information as covariates. We also consider spatial dependency among the monitoring sites to provide a more flexible modelling framework and increase interpretability. Our goal is to model the number of days per month overcoming two different threshold criteria: the *single-hour threshold*, i.e. at least one hour (in a day) with average ozone concentration over $180 \mu\text{g}/\text{m}^3$, and the *sliding-window threshold*, i.e. at least one running average over eight consecutive hours (in a day) greater than $120 \mu\text{g}/\text{m}^3$. We then aggregate information to obtain the two corresponding datasets. The single-hour threshold allows us to detect

extremely high spikes of ozone concentrations, while the sliding-window threshold provides information about the overall dangerous level of O_3 pollution during the day. We aim to identify and interpret any difference between these two over-threshold counts to better understand the phenomenon of ozone pollution in its entirety. In Sect. 2, we describe the ozone datasets and the covariates included in our analysis. Section 3 illustrates the Bayesian model we propose, and Sect. 4 reports corresponding posterior inference.

2 The Ozone Dataset

The original dataset consists of the hourly ozone (O_3) concentrations recorded by *ARPA Lombardia* monitoring network. These data were gathered from 2010 to 2022 by 45 monitoring stations across Lombardy (Northern Italy). Following experts' suggestions, we focus only on a specific time of the year, from April to October, when ozone levels are notoriously higher. Other months present a negligible number of days over the thresholds, becoming useless for the purpose of our analysis. As the raw hourly data presented many missing values, informed decisions were taken when constructing the two count datasets. We impose reasonable constraints to deal with not available (NA) observations when summarising the data in terms of monthly counts. For the construction of the *single-hour threshold* dataset, it is important to understand that: (i) a day is labelled as NA if it lacks six or more hourly data and the available concentrations do not overcome the threshold of $180 \mu\text{g}/\text{m}^3$, (ii) a month with six or more missing days is considered non-informative, and hence labelled as NA in the final dataset. At this point, we have around 10% of months with at least one missing day, most of which happen to have more than five missing days, being labelled as NAs and modelled accordingly into our Bayesian framework. For the few months with five or fewer missing days, the daily value is simulated using a linear interpolation using the nearest preceding and succeeding observed days so that we do not have to discard the whole month from the observed data.

When considering the *sliding-window threshold*, dealing with missing values becomes even more challenging, as we must consider averages over multiple consecutive hours for each day. Here, we assume the overall average value is included in our dataset if there are at least six observed concentrations over the eight-hour horizon. Once the dataset of rolling averages is generated, the imputation of daily counts exceeding the $120 \mu\text{g}/\text{m}^3$ threshold is the same used in the *single-hour threshold* scenario. Finally, we apply the same criteria as above to label a whole day or month as NA and update the monthly counts accordingly.

Summing up, we consider two datasets with monthly counts of days presenting high ozone levels above two different thresholds for each year and each monitoring station. We aim to investigate if these two variables, and the models representing them, highlight different aspects of O_3 pollution. Specifically, we are interested in studying the evolution of time-dependent effects and the influence of meteorological covariates in the two scenarios.

2.1 Covariates

It is well-known that meteorological factors strongly affect the incidence of air pollutants in the atmosphere. For each location, we get daily values of meteorological factors through **OpenMeteo** [6] historical archives. Daily information is then aggregated into monthly summaries to get mean temperature ($^{\circ}\text{C}$), total sum of precipitations (mm), hours with precipitations (h), maximum wind speed (km/h) and total sum of shortwave radiations (MJ/m^2). We then use the monitoring stations' geographical location (longitude and latitude) to define a proper spatial covariance structure for our data.

3 The Binomial Regression Model

When dealing with count data in a regression context, one typically defaults to Poisson or negative binomial generalized regression models (see for example [2, 5]). However, these models are best fitted for count data with an unknown or non-existent upper limit. In this study instead we model over-threshold days per month, so we must account for the natural upper limit given by the total number of days in each month. Given this limitation, our specific scenario is best suited for a binomial regression model. For both datasets, we model the number of days in a month overcoming a given threshold. We denote this count by n_{ijk} for each month $i = 1, \dots, 7$ (from April to October), year $j = 1, \dots, 13$ (from 2010 to 2022), and monitoring station $k = 1, \dots, 45$. Here $N_{apr}, \dots, N_{oct}$, the total number of days in each month, are of course assumed to be fixed. The observation-specific over-threshold probabilities $\theta_{i,j,k}$ are then characterized through a generalized linear model with a hierarchical structure. In fact, using the logit link function, we assume:

$$\begin{aligned} n_{ijk} | \theta_{ijk} &\overset{ind}{\sim} \text{Bin}(N_i, \theta_{ijk}), \quad i = 1, \dots, 7, \quad j = 1, \dots, 13, \quad k = 1, \dots, 45 \\ \text{logit}(\theta_{ijk}) &= \mathbf{x}_{ijk}^T \boldsymbol{\beta} + \gamma_i + \xi_j + w_k, \end{aligned} \quad (1)$$

where $\mathbf{x}_{ijk}$ is the p -dimensional vector of standardized meteorological covariates presented in Sect. 2.1, for each month i , year j and station k . Equation (1) defines a mixed-effects model where ξ_j is the annual effect, γ_i is the monthly effect, and w_k is the station-specific effect. While annual and monthly parameters will be treated as conditional independent random effects, the station-specific term will be modeled as a spatial Gaussian process (GP) in order to account for spatial correlation. A priori $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\xi}$ and $\mathbf{w}$ are independent, with marginal priors described here below:

$$\gamma_i \mid \sigma_{\gamma}^2 \overset{iid}{\sim} \mathcal{N}(\mu_{\gamma}, \sigma_{\gamma}^2) \text{ for } i = 1, \dots, 7 \quad (2)$$

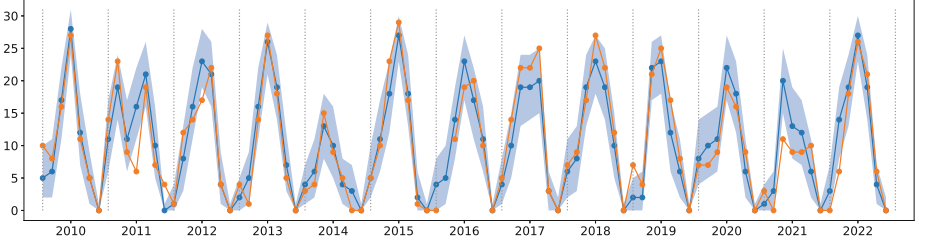
$$\xi_j \mid \sigma_{\xi}^2 \overset{iid}{\sim} \mathcal{N}(\mu_{\xi}, \sigma_{\xi}^2) \text{ for } j = 1, \dots, 13. \quad (3)$$

We do not assume here any explicit temporal dependency, since from exploratory data analysis we have seen that the only clear temporal pattern

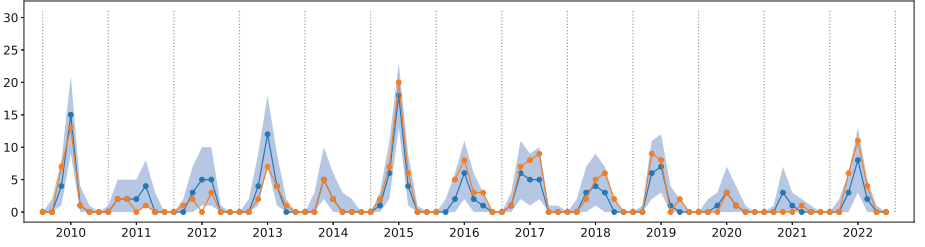
is the presence of higher counts in hotter months, as one could expect. This specific behaviour is already explained by γ , without the need to include a more complex time-dependent structure. The marginal prior of $\mathbf{w} = (w_1, \dots, w_{45})$ is characterized by an exponential covariance function depending on the distance between site locations (see [4]), i.e., $H_{ij} = \exp\left(\frac{1}{\rho} \text{dist}(\mathbf{s}_i, \mathbf{s}_j)\right)$ with $H = [H_{i,j}]$ and

$$\mathbf{w} \mid \Sigma, \sigma_s^2 \sim \mathcal{N}(\mathbf{0}, \Sigma + \sigma_s^2 H) \quad (4)$$

where ρ is the spatial range and $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_{45}^2)$ accounts for station-specific nugget effects. We assign a Gaussian marginal prior distribution to



(a) Sliding-window threshold



(b) Single-hour threshold

Fig. 1. 95% CI (light-blue) and median (blue) of the posterior predictive distribution for the over-threshold O_3 at station 10081 with observed values (orange). (Color figure online)

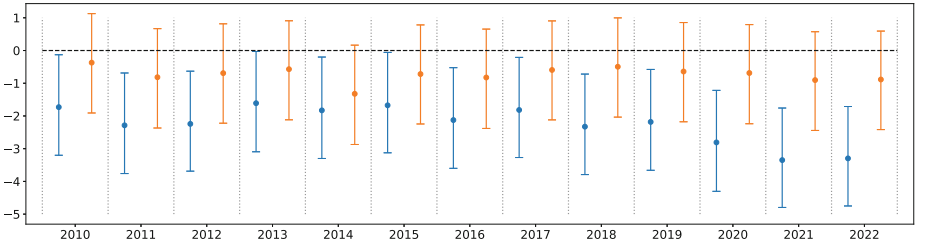


Fig. 2. Marginal posterior 95% CIs for year-specific effects ξ considering single-hour (blue) or sliding-window (orange) threshold. (Color figure online)

the regression coefficients $\beta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We complete the model with suitable marginal priors for hyperparameters in (2)–(3), i.e., following [3] we assume

$$\mu_\gamma, \mu_\xi \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad \sigma_\gamma^2, \sigma_\xi^2, \sigma_s^2, \sigma_1^2, \dots, \sigma_{45}^2 \stackrel{iid}{\sim} \text{Inv-Gamma}(4, 2). \quad (5)$$

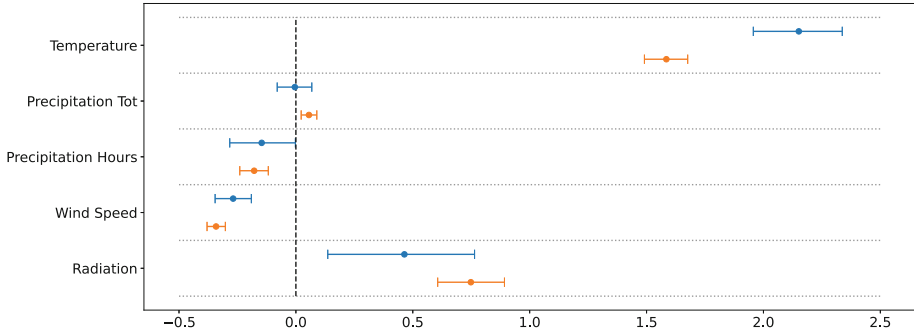


Fig. 3. Marginal posterior 95% CIs for regression coefficients β considering single-hour (blue) or sliding-window (orange) threshold. (Color figure online)

4 Posterior Inference

We apply model (1)–(5) to the two over-the-threshold counts. Posterior inference provided in this section has been obtained running the MCMC in *Stan* [1] assuming 4 parallel chains, each of them with burn-in = 1000 and 1000 sampling iterations, for a final sample size of 4000 iterations. Figure 1 shows the median and 95% credible intervals (CI) of the posterior predictive distribution for the target variable at one of the monitoring sites. Our prediction follows closely the observed data. To assess the predictive goodness-of-fit of our model, we randomly select 30% of the observations to use as a test set, while the other 70% is used to train the model. We observe that 96% (87%) of observed values are actually covered by the 95% CI of the posterior predictive distribution for the single-hour (sliding-window) over-threshold counts. Lower coverage in the sliding-window scenario might be explained by the underestimation of upper extreme values and overestimation of lower ones, as can be seen in Fig. 1a.

We focus now on the random effects and regression coefficients. Looking at the marginal posterior distributions of time-varying parameters, the year-specific parameters ξ of the two datasets seem to follow different trends, while the monthly random effects γ do not differ and hence are not reported here. Figure 2 displays the marginal posterior 95% credible intervals for ξ , comparing the posterior estimates for the single-hour (blue) and sliding-window (orange) thresholds. In the sliding-window scenario we see no specific pattern, while with the single-hour threshold we identify a decreasing trend. This behaviour can be interpreted

as a decrease in the annual effect of extreme events ($O_3 > 180$) while the overall level (sliding-window) of ozone pollution does not decrease over the years. Considering instead the marginal posterior 95% CIs for the regression coefficients β in Fig. 3, we see that the meteorological factors similarly affect the responses in the two datasets. Temperature and solar radiation increase the over-threshold probability, while wind and precipitation are associated with lower values. These estimates are coherent with pollution studies. The plot of the marginal posterior 95% CIs for the estimated spatial residuals w in the two scenarios, not reported here, shows that there are clear differences between the two responses, and from station to station. Further details about the datasets, MCMC codes and posterior inference are available at https://github.com/venanzjones/Bayesian_Project.

Acknowledgments. We thank the students of *Bayesian statistics* course (MSc in Mathematical Engineering of Politecnico di Milano) A. Ganelli, F. M. Mancinelli, B. Sisti and A. Venanzi for their contribution in implementing the Stan code. Michela Frigeri and Alessandra Guglielmi acknowledge the support of MUR, Grant Dipartimento di Eccellenza 2023–2027.

References

1. Carpenter, B., et al.: Stan: a probabilistic programming language. *J. Stat. Softw.* **76**(1) (2017)
2. Davis, R.A., et al.: Count time series: a methodological review. *J. Am. Stat. Assoc.* **116**(535), 1533–1547 (2021)
3. Gelman, A.: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1**(3), 515–534 (2006)
4. Sahu, S.K.: *Bayesian Modeling of Spatio-Temporal Data with R*. Chapman and Hall/CRC (2022). <https://doi.org/10.1201/9780429318443>
5. Sellers, K.F., Shmueli, G.: A flexible regression model for count data. *Ann. Appl. Stat.* **4**(2), 943–961 (2010)
6. Zippenfenig, P.: Weather API (2023). <https://open-meteo.com/>



Adaptive Elastic-Net Estimation for Ergodic Diffusion Processes

Dario Frisardi^{1,2(✉)}, Alessandro De Gregorio¹, Francesco Iafrate¹,
and Stefano M. Iacus³

¹ Sapienza University of Rome, Rome, Italy

{dario.frisardi,alessandro.degregorio,francesco.iafrate}@uniroma1.it

² Italian National Institute of Statistics - Istat, Rome, Italy

³ Harvard University, Cambridge, MA, USA

siacus@iq.harvard.edu

Abstract. This paper introduces the Adaptive Elastic-net estimator for ergodic diffusion processes, demonstrating oracle properties in an asymptotic regime with mixed rates, crucial for such processes, ensuring consistent selection of relevant variables and optimal convergence rates. Additionally, it describes an iterative algorithm for parameter estimation, along with presenting the regularization path in a simulation study. This paper serves as a starting point for further investigation into various types of stochastic processes, offering potential improvements in prediction accuracy and interpretability.

Keywords: regularized estimator · oracle · optimization algorithm

1 Introduction to Adaptive Elastic-Net

In recent years, penalized statistical learning has become increasingly crucial, underscoring the significance of accurate prediction and the discovery of relevant predictive variables, especially in scenarios where sparse representations are inherent in the underlying model [6, 7].

The Elastic-net is a regularization technique which combines the penalties of both the Lasso (ℓ_1 -penalty) and Ridge (ℓ_2 -penalty) methods and simultaneously performs estimation and variable selection. This penalization offers a robust solution, particularly when dealing with multicollinearity and high-dimensional feature space. As a matter of fact, it overcomes several limitations of Lasso, such as selecting at most n components, if $p > n$, and detecting grouping information in the case of highly correlated pairwise variable groups [14].

The Adaptive Elastic-net penalization is proposed, using a combination of the ℓ_2 -penalty and the adaptive ℓ_1 -penalty to ensure the consistency of variable selection, by assigning a different amount of shrinkage to each component. Thus, the Adaptive Elastic-net estimator, as defined in [15], is:

$$\hat{\theta}_{AdaEnet} \in \arg \min_{\theta} \left\{ \mathcal{L}_n(\theta) + \lambda \sum_{j=1}^n |\theta_j|^2 + \sum_{j=1}^p \hat{w}_j |\theta_j| \right\} \quad (1)$$

where $\mathcal{L}_n(\theta)$ is a generic loss function, $\lambda \geq 0$ is a shrinking parameter and $(\hat{w}_j)_{j=1}^p$ are data-driven adaptive weights. If we force λ to be zero in (1), then the Adaptive Elastic-net reduces to Adaptive Lasso.

A good selection procedure should have (asymptotically) the so-called oracle properties [5]:

- i) the selection procedure identifies the right subset model;
- ii) has the optimal convergence rate and converges to a Gaussian random variable $N(0, \Sigma_0)$ where Σ_0 is the covariance matrix of the true subset model.

The Adaptive Lasso fixes the consistency selection issue of the Lasso procedure in certain scenarios, thus adhering to the oracle properties [13]. The Adaptive Elastic-net penalization outperforms both the Adaptive Lasso and the Elastic-net in a high-dimensional framework [15].

Originally developed for linear regression problems, procedures with sparse representation have been increasingly applied across fields of stochastic processes [10, 11] and multidimensional diffusion processes [3, 4, 8, 9] in recent years.

In this short paper, we introduce the Adaptive Elastic-net estimator within the framework of ergodic diffusion processes, where a mixed-rates asymptotic regime must be considered. In Sect. 2, we introduce the notation for a diffusion process and discuss the asymptotic properties of the Adaptive Elastic-net estimator with varying rates of convergence. Section 3 presents the explicit elementwise solution in an optimization algorithm, along with the estimation of coefficient paths in a simulation study. The final section outlines forthcoming developments aimed at improving efficiency and prediction accuracy.

2 Regularized Estimator for Ergodic Diffusion Processes

2.1 Ergodic Diffusion Processes

Let $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t)_{t \geq 0}, P)$ be a filtered complete probability space. Let us consider a d -dimensional solution process $X := (X_t)_{t \geq 0}$ to the following stochastic differential equation (SDE):

$$dX_t = b(X_t, \alpha)dt + \sigma(X_t, \beta)dW_t, \quad X_0 = x_0 \quad (2)$$

where x_0 is a deterministic initial point, the drift coefficient $b : \mathbb{R}^d \times \Theta_\alpha \rightarrow \mathbb{R}^d$ and the diffusion coefficient $\sigma : \mathbb{R}^d \times \Theta_\beta \rightarrow \mathbb{R}^d \otimes \mathbb{R}^r$ are Borel known functions, up to α and β , and $(W_t)_{t \geq 0}$ is a r -dimensional standard $\mathcal{F}_t$ -Brownian motion. Moreover, $\alpha \in \Theta_\alpha \subset \mathbb{R}^p$ and $\beta \in \Theta_\beta \subset \mathbb{R}^q$, with $p, q \in \mathbb{N}$, are unknown parameters where $\Theta_\alpha, \Theta_\beta$ are compact convex sets.

X satisfies some mild conditions on b and σ (see, e.g., [12] for a discussion on the model assumptions) and it is an ergodic diffusion. The sample path of X is observed at $n+1$ equidistant discrete times t_i^n , such that $t_i^n - t_{i-1}^n = \Delta_n < \infty$ for $i = 1, \dots, n$ with $t_0^n = 0$. Data are the discrete observations of the sample path of X , that we represent by $\mathbf{X}_n := (X_{t_i^n})_{0 \leq t_i \leq n}$.

The asymptotic scheme adopted is $n\Delta_n \rightarrow \infty$ and $\Delta_n \rightarrow 0$, as $n \rightarrow \infty$, and $\exists \epsilon \in (0, \frac{e-1}{e})$, with $e \in \mathbb{N}_{\geq 2}$, such that $n \leq n\Delta_n$, for large n .

The parameter of interest is $\theta := (\alpha, \beta)^\top \in \Theta := \Theta_\alpha \times \Theta_\beta \subset \mathbb{R}^{p+q}$. Let us denote $\theta_0 := (\alpha_0, \beta_0)$ the true value of θ and assume that $\theta_0 \in \text{Int}(\Theta)$ and $0 \in \mathbb{R}^{p+q}$ belongs to Θ . The SDE X represents a sparse parametric model, that is θ_0 has a sparse representation, with $\mathbf{p}^0 := |\{j : \alpha_{0,j} \neq 0\}|$ and $\mathbf{q}^0 = |\{h : \beta_{0,h} \neq 0\}|$. Precisely, let us assume $\alpha_{0,j} \neq 0$ for $j = 1, \dots, \mathbf{p}_0$ and $\beta_{0,h} \neq 0$ for $h = 1, \dots, \mathbf{q}_0$. Moreover, we denote $\alpha_\star := (\alpha_1, \dots, \alpha_{\mathbf{p}^0})^\top$, $\beta_\star := (\beta_1, \dots, \beta_{\mathbf{q}^0})^\top$ and $\alpha_\bullet := (\alpha_{\mathbf{p}^0+1}, \dots, \alpha_{\mathbf{p}})^\top$, $\beta_\bullet := (\beta_{\mathbf{q}^0+1}, \dots, \beta_{\mathbf{q}})^\top$.

2.2 Adaptive Elastic-Net Estimator and Oracle Properties

The aim of this paper is to introduce the Adaptive Elastic-net estimator for discretely observed ergodic diffusion.

Let $\mathcal{L}_n(\theta)$ be a generic loss function for θ based on the sample $\mathbf{X}_n$ (for instance the quasi-likelihood function adopted in [12]). Let us consider an initial estimator $\tilde{\theta}_n$ minimizing $\mathcal{L}_n(\theta)$, twice differentiable with respect to θ .

The initial estimator $\tilde{\theta}_n$ satisfies the following asymptotic properties:

P1) $\tilde{\theta}_n$ is consistent, i.e.,

$$\left(\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0), \sqrt{n}(\tilde{\beta}_n - \beta_0) \right)^\top = O_p(1);$$

P2) $\tilde{\theta}_n$ is asymptotically normal, i.e.

$$\left((\sqrt{n\Delta_n}(\tilde{\alpha}_n - \alpha_0), \sqrt{n}(\tilde{\beta}_n - \beta_0))^\top \xrightarrow{d} N_{\mathbf{p}+\mathbf{q}}(0, \Gamma^{-1}) \right)$$

where Γ is a $(\mathbf{p} + \mathbf{q}) \times (\mathbf{p} + \mathbf{q})$ positive definite symmetric matrix.

Thus from P1 and P2, $\tilde{\theta}_n$ works in a mixed-rates asymptotic regime with two different rates, $\sqrt{n\Delta_n}$ for α and $\sqrt{n}$ for β .

The objective function is obtained from Least Square Approximation (LSA) (see [10]) of $\mathcal{L}_n(\theta)$:

$$\mathcal{F}_n(\theta) = (\theta - \tilde{\theta}_n)^\top \hat{G}_n(\theta - \tilde{\theta}_n) + R(\theta) + L(\theta) \quad (3)$$

where $\hat{G}_n$ is a $(\mathbf{p} + \mathbf{q}) \times (\mathbf{p} + \mathbf{q})$ almost surely positive definite symmetric random matrix depending on n . Notably, $R(\theta)$ represents the Ridge component (ℓ_2 -penalty):

$$R(\theta) = \psi_n \sum_{j=1}^{\mathbf{p}} |\alpha_j|^2 + \gamma_n \sum_{h=1}^{\mathbf{q}} |\beta_h|^2 \quad (4)$$

where ψ_n and γ_n are sequence of positive numbers, depending on n . The Adaptive Lasso component (adaptive ℓ_1 -penalty):

$$L(\theta) = \sum_{j=1}^{\mathbf{p}} \kappa_{n,j} |\alpha_j| + \sum_{h=1}^{\mathbf{q}} \pi_{n,h} |\beta_h| \quad (5)$$

where $\kappa_{n,j}$ and $\pi_{n,h}$ are sequence of real positive random variable representing the amount of shrinkage for α_j and β_h , $j = 1, \dots, p$ and $h = 1, \dots, q$, respectively.

Furthermore, let us consider $A_n \hat{G}_n A_n \xrightarrow{p} G$ where $A_n := \text{diag} \left(\frac{1}{\sqrt{n}\Delta_n} \mathbf{I}_p, \frac{1}{\sqrt{n}} \mathbf{I}_q \right)$, with $\mathbf{I}_g$ is the identity matrix of size g , and G is a positive definite symmetric random matrix (see [4, 10]).

The main idea is to deal with ℓ_1 and ℓ_2 penalties for each group of parameters.

Definition 1. *The Adaptive Elastic-net estimator $\hat{\theta}_n : \mathbb{R}^{(n+1) \times d} \rightarrow \bar{\Theta}$ is defined*

$$\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)^\top \in \arg \min_{\theta \in \bar{\Theta}} \mathcal{F}_n(\theta) \quad (6)$$

From the properties above and some assumptions of convergence, for tuning parameters and adaptive weight, we demonstrate the following theorems, showing that the Adaptive Elastic-net estimator (6) satisfies the oracle properties in an asymptotic regime with mixed rates, as in the context of ergodic diffusion processes.

Theorem 1 (Consistency). *From property P1, then*

$$A_n^{-1} (\hat{\theta}_n - \theta_0) = O_p(1)$$

Theorem 2 (Selection consistency). *From property P1, we have that:*

$$P(\hat{\alpha}_{n\bullet} = 0) \longrightarrow 1 \text{ and } P(\hat{\beta}_{n\bullet} = 0) \longrightarrow 1$$

as $n \longrightarrow \infty$. Then, $P(\hat{\theta}_{n\bullet} = 0) \longrightarrow 1$.

Theorem 3 (Asymptotic normality). *From properties P1 and P2, then*

$$\left(\sqrt{n\Delta_n}(\hat{\alpha}_n - \tilde{\alpha}_n), \sqrt{n}(\hat{\beta}_n - \tilde{\beta}_n) \right)^\top \xrightarrow{d} N_{p^0+q^0}(0, \Gamma_\star^{-1})$$

as $n \longrightarrow \infty$.

3 Optimization Algorithm and Coefficients Path

To obtain an explicit solution in an optimization problem where the objective function can be decomposed as $f = g + h$, with g being convex and differentiable, and h being convex but not differentiable, iterative proximal gradient algorithms can be used. Specifically, accelerated gradient methods use a weighted combination of the current estimate and the previous gradient directions, and the proximal map function is iteratively applied till convergence of the algorithm (see [1, 7] for further information on first-order algorithms).

In particular, in the objective function (3) we can consider the adaptive ℓ_1 -penalty as the function h in the above decomposition. It is possible to use the coordinate descent iterative algorithm where, at iteration t , the update is:

$$\hat{\theta}_k^{t+1} = \arg \min_{\theta_k} \mathcal{F}_n \left(\hat{\theta}_1^t, \dots, \hat{\theta}_{k-1}^t, \theta_k, \hat{\theta}_{k+1}^t, \dots, \hat{\theta}_{p+q}^t \right) \quad (7)$$

The explicit elementwise update, for $k = 1, \dots, \mathbf{p} + \mathbf{q}$ and $t \geq 0$, is:

$$\hat{\theta}_k^{t+1} = \frac{1}{1 + \frac{\lambda}{\hat{G}_{(k,k)}}(1 - \xi)} \mathcal{S}_{\frac{\lambda\xi}{\hat{G}_{(k,k)}}w_k} \left(\tilde{\theta}_k - \frac{1}{\hat{G}_{(k,k)}} \sum_{i \neq k} \hat{G}_{(i,k)} (\hat{\theta}_k^t - \tilde{\theta}_k) \right) \quad (8)$$

where λ is the tuning parameter in the ℓ_2 -penalty, $\hat{G}_{(i,k)}$ is the element in the i -th row and k -th column of $\hat{G}_n$, $\xi \in (0, 1]$ controls the relative proportion between ℓ_1 and ℓ_2 regularization penalties (i.e. $\xi = 1$, it reduces to Adaptive Lasso), w_k is the adaptive weight for element k in the ℓ_1 -penalty, and $\mathcal{S}_\mu(z) := \text{sign}(z)(z - \mu)_+$ is the soft-thresholding operator.

Improvements beyond the reported algorithm can be achieved by leveraging accelerated variants such as the Fast Iterative Soft-Thresholding Algorithm (FISTA, see [1]) and Proximal Accelerated Gradient Descent Algorithm [7], along with a blockwise Proximal Alternating Minimization (referred to in literature as PALM, see [2]), which updates one block of parameters at each step.

In a diffusion process, as in (2), the classical cross-validation technique for tuning parameter selection is not applicable due to the data dependency structure. For this reason, a data-driven technique is used, based on an iterative algorithm that considers the Euler discretization of the solution of (2) and, given estimates of the parameters, a score function of the current estimated residuals (see [4]).

We consider a multivariate Ornstein-Uhlenbeck diffusion process, which can be written using the following notation:

$$dX_t = -BX_t dt + AdW_t \quad (9)$$

where B and A are SPD matrices in $\mathbb{R}^{d \times d}$, containing the drift and diffusion parameters. In Fig. 1 the coefficients paths from model (9) with $d = 5$.

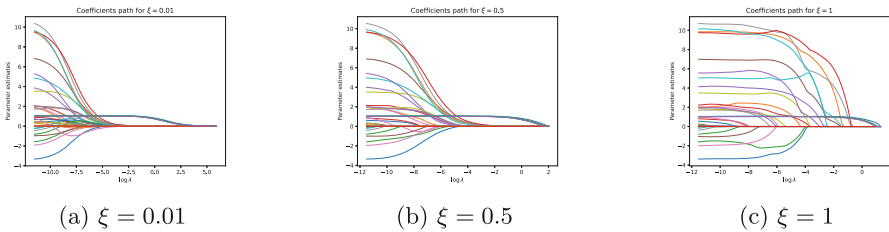


Fig. 1. Adaptive Elastic-net coefficients paths for different values of the proportion parameter ξ . In (c), $\xi = 1$, it coincides with Adaptive Lasso solutions.

4 Further Developments

In our forthcoming investigations, we will conduct a comprehensive evaluation of the efficiency of the Adaptive Elastic-net estimator, focusing on its effectiveness

in managing mixed-rate regimes for both estimation and prediction tasks. Our study will involve applying this approach to real-world data, aiming to assess its practicality and effectiveness in applied contexts. By extending our research beyond ergodic processes, we aim to include a wide range of stochastic processes to broaden the applicability of this approach across diverse contexts, especially within the realm of high-dimensional data analysis.

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.* **2**(1), 183–202 (2009)
2. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.* **146**(1–2), 459–494 (2014)
3. De Gregorio, A., Iacus, S.M.: Adaptive LASSO-type estimation for multivariate diffusion processes. *Economet. Theor.* **28**(4), 838–860 (2012)
4. De Gregorio, A., Iafrate, F.: Regularized bridge-type estimation with multiple penalties. *Ann. Inst. Stat. Math.* **73**(5), 921–951 (2021)
5. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001)
6. Fu, W., Knight, K.: Asymptotics for lasso-type estimators. *Ann. Stat.* **28**(5), 1356–1378 (2000)
7. Hastie, T., Tibshirani, R., Wainwright, M.: *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press (2015)
8. Kinoshita, Y., Yoshida, N.: Penalized quasi likelihood estimation for variable selection (2019). arXiv preprint [arXiv:1910.12871](https://arxiv.org/abs/1910.12871)
9. Masuda, H., Shimizu, Y.: Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. *Math. Methods Statist.* **26**(2), 81–110 (2017). <https://doi.org/10.3103/S1066530717020016>
10. Suzuki, T., Yoshida, N.: Penalized least squares approximation methods and their applications to stochastic processes. *Jpn J. Stat. Data Sci.* **3**(2), 513–541 (2020). <https://doi.org/10.1007/s42081-019-00064-w>
11. Wang, H., Leng, C.: Unified LASSO estimation by least squares approximation. *J. Am. Stat. Assoc.* **102**(479), 1039–1048 (2007)
12. Yoshida, N.: Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Stat. Math.* **63**(3), 431–479 (2011)
13. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)
14. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat Methodol.* **67**(2), 301–320 (2005)
15. Zou, H., Zhang, H.H.: On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **37**(4), 1733 (2009)



Internet Activities and Life Satisfaction Among Older People in Italy

Claudia Furlan^(✉) and Silvia Meggiolaro

Department of Statistical Sciences, University of Padova, Padova, Italy
claudia.furlan@unipd.it, meg@stat.unipd.it

Abstract. The study aims to provide new insights regarding the role of Internet use for the wellbeing of older people in Italy, taking into account the different activities performed and potential differences according to the living arrangements of older people and their gender. Data come from the survey “Aspects of Daily Life”, conducted by the Italian National Institute of Statistics (Istat) in 2021. Results of multinomial logistic regression models suggest that both for older men and women, and independently from the living arrangement, the use of Internet can have a role for the life satisfaction, but Internet use presents associations with life satisfaction that are differentiated according to the type of activities performed using Internet, and according to gender and living arrangement. In particular, men living alone seem to benefit from calls and video-calls, instead, men living with others have higher life satisfaction if they use Internet for texting. For women living alone there are not significant associations between life satisfaction and activities performed using Internet use; however, women living with others present a positive relationship between high levels of life satisfaction and video-calls.

Keywords: Internet Use · Life satisfaction · older people

1 Introduction

With the progressive ageing of societies, increasing attention in social research and among policy makers has been directed towards older adults and the ageing process. Two particular areas of interest are the so-called “successful ageing” and older people’s wellbeing. The term “successful ageing” was popularized by Rowe and Kahn [1], who defined the term as including three main components: “low probability of disease and disease-related disability, high cognitive and physical functional capacity, and active engagement with life” [2]. Older people’s “wellbeing,” meanwhile, is said to be the subjective counterpart of successful ageing [3], and this explains the attention that the wellbeing of older people receives in literature as well as in policy-making processes. Specifically, among the different aspects of subjective wellbeing that can be distinguished, the evaluative dimension measured by life satisfaction (LS) is usually considered a key indicator of wellbeing (see, for example, [4, 5]) and is one of the most commonly used measures of successful aging too [6]. This explains the interest in life satisfaction by social research.

Previous studies have suggested that subjective wellbeing (SWB) is influenced by many factors and, in particular, it is positively associated both with social contact and social integration [7, 8] and with being active [9]. However, both social networks and active lifestyles in later life decrease due to retirement, potential health problems and death of family member and friends. As a consequence, social isolation and loneliness can increase, leading to a deterioration of wellbeing; similarly, the sense of usefulness and the mental stimulation decrease when older adults are less active, implying, again, a decline in wellbeing.

In recent years, Internet use has become an important tool for social activities and leisure time among people, and it may be important among older people as well to reduce both loneliness and the loss of meaning of their lives, increasing their wellbeing [10]. It can be, indeed, associated with social involvement in new personal friendships, and with family and friends [11].

The aim of the current study is to provide new insights regarding the role of the use of Internet for older people's wellbeing, distinguishing not between users and not users, but also considering the specific activities (connected with communication and social network integration) carried out in Internet, and taking into account the potential differences according to the living arrangements of older people and their gender.

The reference is a country such as Italy, which represents an interesting case study for various reasons. Italy's population is ageing at an unprecedented rate, with a rapid rise in the numbers of older people, along with persistently low fertility and high life expectancy [12]. Italy also offers an interesting context to study the gender differences, being characterized by a still relatively unbalanced gender system [13], and thus differences in socialisation of old men and women could be particularly strong in shaping expectation and behaviour in later life (and thus, SWB). Lastly, in Italy the percentage of individuals using the Internet is below the EU average (particularly among older people),¹ suggesting that if Internet has a positive role on SWB, this aspect has to receive further attention to possibly increase this percentage.

2 Data and Methods

We use data from the survey "Aspects of Daily Life", undertaken in Italy in 2021 by the Italian National Institute of Statistics (Istat); it is based on a nationally representative sample of approximately 24,000 households, with a total of 45,607 individuals. In our study, we focused on 12,299² individuals aged 65 and above. The survey collected information on several dimensions of life. Besides socio-demographic and socio-economic characteristics, information on health, lifestyles, religious practices, and social integration for each household member were recorded. In addition, a specific section in the questionnaire investigates the Internet use.

As regards the focus of our study, life satisfaction was assessed with the question: "How satisfied are you with your life on the whole at present?", with answers ranging from 0 (not satisfied at all) to 10 (very satisfied). The main explanatory variables

¹ https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ifp_iu/default/table?lang=en.

² 280 individuals (2.2%) were excluded from the analyses due to missing values for our key variable that is life satisfaction.

of interest for our study are those describing the use of Internet. In particular, the survey investigates, for those individuals who declared they have used Internet in the last three months, the types of activities performed. Specifically, we considered 5 activities connected with communication and social network integration:

1. sent or received e-mails;
2. called or video-called through internet (as with Skype, Messenger, WhatsApp, Facetime, Viber, Snapchat);
3. used instant message services (as WhatsApp, Skype, Viber, Snapchat);
4. used a social network (as Twitter, Facebook, Instagram, Snapchat);
5. expressed opinions on social or political issues on social networks or taken part in on-line consultation/voting on civic/social/political issues (for example, urban planning, to sign a petition).

We estimated multinomial logistic regression models distinguishing three categories of the scores in life satisfaction. Specifically, we considered LS lower than 6, LS equal to 6 or 7, and LS equal or greater than 8.³ Multivariate analyses were applied to study the association between older people's life satisfaction and Internet use, net of the role of socio-demographic characteristics, socio-economic conditions, health status, real social interactions, active lifestyles, and contextual aspects.

Analyses were conducted separately by gender and living arrangement, distinguishing older men and women living alone from those living with other people.

3 Results

Table 1 shows some descriptive statistics of the variable on LS according to gender and living arrangements for our sample of older people: men are, in general, more satisfied than women, and among both men and women, those living alone are less satisfied with life than those living with other individuals.

Table 2 showed that some activities are more common than other ones; however, the proportions of older people using Internet for the various activities are always less than 35%. Confirming previous literature, men are stronger Internet users than women. In addition, if for men the proportions of users are similar according to whether they live alone or with other persons, for women the proportions of users are higher among those living with others.

From Table 3, it is interesting to see that, for men living alone, some signals of benefits of the Internet use can be seen only for an activity connected with communication, that is having had calls and video-calls: this activity is significantly and positively associated with having a medium level of LS, but not with higher levels of LS. This may be explained by the fact that (video) calling is the activity more similar to a real contact and this could be particularly important for those men living alone with a medium level of LS, while the same men having a high level of LS do not benefit from the Internet Use.

³ Preliminary analyses considered alternative categories as robustness checks. The results with a finer distinction of categories (LS lower than 6, LS equal to 6, LS equal to 7, LS equal to 8, and LS greater than 8) were similar to those provided by the more parsimonious models we presented here.

Table 1. Mean and median of LS, and percentage of older people with a LS lower than 6, equal to 6 or 7, and greater than 7, according to gender and living arrangements.

	All individuals		Living alone		Living with others	
LS	Men	Women	Men	Women	Men	Women
Mean	7.21	6.91	7.06	6.79	7.24	6.97
Median	7.00	7.00	7.00	7.00	7.00	7.00
% LS < 6	12.37	17.82	14.42	19.60	11.97	16.86
% LS = 6 - 7	40.04	41.33	43.69	42.57	39.33	40.66
% LS > 7	47.59	40.85	41.89	37.83	48.70	42.48
N	5,432	6,867	4,544	4,459	888	2,408

Table 2. Proportions of older men and women who have used Internet for different activities according to gender and living arrangement.

	Living alone		Living with others	
Internet Use	Men	Women	Men	Women
1. e-mail	28.8	13.4	29.2	16.4
2. Call and video call	24.8	16.7	26.5	23.1
3. Messaging app	29.0	19.9	32.5	27.3
4. Social networks	12.5	8.5	12.6	10.5
5. Opinions/voting	8.0	4.7	8.4	5.9

The same activity of (video) calling is not significantly associated with LS for men living with others, for whom another type of communication through Internet has a significant and positive relationship with LS: texting is connected with both medium and high levels of LS. About the activity of texting, results show a synergism effect between Internet use for communication and real social networking that are more implicit living with other persons (in line with results in [14]).

For women living alone, there are not significant associations between life satisfaction and activities performed using Internet use. Instead, some signals of an impact of Internet use on life satisfaction can be found for women living with others. Specifically, a positive association of high levels of LS with (video) calls, the activity more similar to a real contact, emerges: this association is not observed for older women living alone, suggesting again a synergism effect between Internet use for communication and real social networking (as in [14]).

The models control for socio-demographic characteristics, socio-economic conditions, health status, social interactions, active lifestyles, and contextual aspects.

Table 3. Parameter estimates and significance levels of the key covariates on Internet use. Multinomial logistic regression models were separately fitted by gender and living arrangement.

	Men		Women	
Internet Use	LS = 6, 7 (vs LS < 6)	LS > 7 (vs LS < 6)	LS = 6, 7 (vs LS < 6)	LS > 7 (vs LS < 6)
<i>Living Alone</i>				
1. e-mail	−0.892	−0.804	0.059	0.339
2. Call and video call	1.396*	1.014	0.094	0.059
3. Messaging app	0.764	0.878	0.180	0.424
4. Social networks	−0.745	−0.914	0.052	−0.064
5. Opinions/voting	−0.270	0.565	0.015	0.390
<i>Living with others</i>				
1. e-mail	−0.264	−0.349	0.005	−0.037
2. Call and video call	0.013	0.229	0.450	0.611*
3. Messaging app	0.777***	0.520*	−0.235	−0.454
4. Social networks	−0.292	−0.187	−0.16	−0.181
5. Opinions/voting	0.019	−0.043	0.225	0.086

Significance level: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4 Conclusions

Empirical literature on whether Internet use is beneficial, harmful, or indifferent for older adults' wellbeing is still inconclusive, but this study offers preliminary evidence in the Italian context, where Internet use among older people has still a low diffusion, compared with other European countries.

This study is among the first to verify whether the relationship between Internet use and older adults' life satisfaction is differentiated by gender and by living arrangement, taking into account also the different activities performed with Internet; previous studies, indeed, mainly distinguished between users and not users and largely overlooked older people's heterogeneity, with an exception being the work by Li and Zhou [15] considering gender differences in the role of Internet use.

Results of this study offers some implications for policies. This study suggests, indeed, that Internet use may act as potential empowerment for older adults to maintain the connection with other individuals. Therefore, it is important to increase the use of Internet among older adults. However, since this study offers preliminary evidence of a differentiated impact of Internet use by gender and by living arrangement, interventions should take into consideration these differences. In addition, results suggested the importance of distinguishing the different types of activities performed with Internet, and, specifically, that only some activities connected to communication may be particularly important.

References

1. Rowe, J.W., Kahn, R.L.: Human aging: usual and successful. *Science* **237**, 143–149 (1987)
2. Rowe, J.W., Kahn, R.L.: Successful aging. *Gerontologist* **37**(4), 433–440 (1997)
3. Stanley, M., Cheek, J.: Well-being and older people: a review of the literature. *Can. J. Occup. Ther.* **70**(1), 51–59 (2003)
4. Steptoe, A., Dwaton, A., Stone, A.A.: Subjective wellbeing, health, and ageing. *The Lancet* **385**, 640–648 (2015)
5. Dumitrache, C.G., Rubio, L., Rubio-Herrera, R.: Extroversion, social support and life satisfaction in old age: a mediation model. *Aging Ment. Health* **22**(8), 1069–1077 (2018)
6. Tate, R.B., Lah, L., Cuddy, T.E.: Definition of successful aging by elderly Canadian males: the Manitoba follow-up study. *Gerontologist* **43**, 735–744 (2003)
7. Golden, J., Conroyb, R.M., Lawlor, B.A.: Social support network structure in older people: underlying dimensions and association with psychological and physical health. *Psychol. Health Med.* **14**, 280–290 (2009)
8. Meggiolaro, S., Ongaro, F.: Life satisfaction among older people in Italy in a gender approach. *Ageing Soc.* **35**, 1481–1504 (2015)
9. Adams, K.B., Leibbrandt, S., Moon, H.: A critical review of the literature on social and leisure activity and wellbeing in later life. *Ageing Soc.* **31**, 683–712 (2011)
10. Boz, H., Karatas, S.E.: A review on Internet use and quality of life of the elderly. *Cypriot J. Educ. Sci.* **10**(3), 182–191 (2015)
11. Du, P., Wang, B.: How does internet use affect life satisfaction of the Chinese elderly? *Popul. Res.* **44**, 3–17 (2020)
12. Tomassini, C., Lamura, G.: Population ageing in Italy and Southern Europe. In: Uhlenberg, P. (ed.) *International Handbook of Population Aging*. International Handbooks of Population, vol. 1, pp. 69–89. Springer, Dordrecht (2009)
13. Anxo, D., Mencarini, L., Pailhé, A., Solaz, A., Tanturri, M.L., Flood, L.: Gender differences in time use over the life course in France, Italy, Sweden, and the US. *Feminist Econ.* **17**(3), 159–195 (2011)
14. Khalaila, R., Vitman-Schorr, A.: Internet use, social networks, loneliness, and quality of life among adults aged 50 and older: mediating and moderating effects. *Qual. Life Res.* **27**, 479–489 (2018)
15. Li, J., Zhou, X.: Internet use and Chinese older adults' subjective well-being (SWB): the role of parent-child contact and relationship. *Comput. Hum. Behav.* **119**, 106725 (2021)



Integrating Pseudo-time in Network Interaction Analysis to Uncover Genetic Pathways in Thyroid Cancer

Antonino Gagliano^{1(✉)}, Gianluca Sottile¹, Nicolina Sciaraffa²,
Claudia Coronello², and Luigi Augugliaro¹

¹ Department of Economics, Business and Statistics, University of Palermo, Viale
delle Scienze 8, Palermo, Italy
antonino.gagliano02@unipa.it

² Advanced Data Analysis Group, Fondazione Ri.MED, Via Bandiera 11,
Palermo, Italy

Abstract. In the last decade, single-cell RNA sequencing technologies have significantly enhanced our understanding of cell biology. These new technologies have enabled us to explore the intricate network of genetic interactions in thyroid carcinoma. We suggest using a *joint conditional graphical lasso* model to compare tumor tissue and adjacent healthy tissue in a 17-year-old patient. The study's key innovation is the introduction of pseudo-time as a covariate, which enables the isolation of interactions between genes that are specifically associated with tumor pathogenesis. Genes such as *UBE2C*, *TOP2A*, *RRM2*, and *TK1* emerge as central, exhibiting high interconnectedness in tumor tissue.

Keywords: multiple Gaussian graphical models · sparse inference · high-dimensional data · single-cell

1 Introduction

Tumors are composed of various cell types, each with its own transcriptional dysregulation. The introduction of new technologies, such as Single Cell RNA Sequencing (*scRNA-seq*), allows for the capture of gene expression profiles of individual cells, providing a more comprehensive understanding of cancer characterization. Given the heterogeneity of cells, which may be at different stages of differentiation, understanding conditional dependencies between genes is a complex problem.

In this work, we propose using pseudo-time, a measure that can impact the process of cell differentiation, as the exploratory variable in a Joint Conditional Graphical Lasso model (*jcglasso*; [3]). The *jcglasso* model was applied to scRNA data collected from thyroid cancer, one of the most prevalent endocrine carcinomas that is currently the subject of active study and research.

Gene expression was used as the response variable (Y) and pseudo-time was used as the explanatory variable (X), treated through spline bases to account for

the nonlinear dependence structure. This study enables the examination of gene interactions within thyroid cells, both in normal and tumor conditions, while considering cell differentiation. It provides a comprehensive view of the specific molecular dynamics in neoplastic thyroid tissue. The results demonstrate that incorporating pseudo-time enables the capture of disease-specific gene interactions, regardless of cellular differentiation.

2 Dataset Description

This study examines a dataset derived from a single-cell analysis of papillary thyroid carcinoma (PTC) in a 17-year-old female patient. The gene expression levels were measured in two tissue types: primary thyroid tumor (TT) and adjacent normal thyroid tissue (NT) using scRNA-seq technology. The dataset consists of two expression matrices of the same 38349 genes and 9864 cells for primary thyroid tumor and 7997 cells for adjacent normal thyroid tissue. The single-cell RNA-sequencing data is available at Accession Number GEO: GSE241184 [1].

3 Methods

The proposed methodology involves two main steps. Firstly, calculating pseudo-time using Tools for Single Cell Analysis (*TSCAN*; [2]). Secondly, using the *jcglasso* model to study gene interactions while accounting for the effect of cell differentiation determined by pseudo-time.

3.1 TSCAN Pseudo-time

TSCAN is structured in three key steps. The first step is to group cells into clusters, the second step is to construct the Minimum Spanning Tree (MST) to connect the centers of each cluster and to establish their order in the differentiation process, and finally, the third step is to project cells along the tree structure to establish their pseudo-temporal value [2].

3.2 Joint Conditional Graphical Lasso

The model presented in [3] expands on the methodology introduced in [4], allowing for the analysis of data with missing-at-random and censored observations. Given $\{\mathcal{T}\} = \{\{B\}, \{\Theta\}\}$ the set of model parameters, in our model, $\{B\} = B_{1,\dots,K}$ represents the spline basis matrix for pseudo-time covariate, and $\{\Theta\} = \Theta_{1,\dots,K}$ is the matrix of partial correlation coefficients, representing sparse gene interaction, for each of the K conditions. This is a crucial element of our analysis as it offers distinct insights for each tissue type (NT and TT), revealing significant differences in characterizing TT genes. The estimator function is:

$$\{\hat{\mathcal{T}}\} = \arg \max_{\{\mathcal{T}\}} \bar{\ell}(\{\mathcal{T}\}) - \rho \tilde{P}_{\alpha_2}(\{\Theta\}). \quad (1.1)$$

where $\bar{\ell}(\{\Upsilon\})$ represents the average log-likelihood function of a multivariate Gaussian distribution, ρ is the tuning parameter for gene interactions (in $\{\Theta\}$) and $\tilde{P}_{\alpha_2}$ has the following specification

$$\tilde{P}_{\alpha_2}(\{\Theta\}) = \alpha_2 \sum_{k=1}^K \sum_{h \neq m} |\theta_{hm,k}| + (1 - \alpha_2) \sum_{h \neq m} \left(\sum_{k=1}^K \theta_{hm,k}^2 \right)^{1/2}, \quad (1.2)$$

where $k = 1, \dots, K$, represent the two distinct conditions (NT and TT) and $\alpha_2 \in (0, 1)$ controls the trade-off between a Lasso penalty, applied to individual pairs of genes, and a Group Lasso penalty, used to encourage shared pattern of sparsity for groups of genes.

4 Results

The analysis is conducted in two main steps. Firstly, we will determine the pseudo-time per single cell. Secondly, we will use the *jcglasso* model to jointly estimate the genetic interactions in primary thyroid tumor (TT) and adjacent normal thyroid tissue (NT), while accounting for the dependence on the differentiation process marked by pseudo-time.

Step 1. After a thorough data preparation phase that included quality checks, normalization, and selection of the most variable genes, we performed dimensionality reduction using PCA and UMAP to cluster and annotate cells in the data. The identified cell types include *Endothelial cells*, *Neutrophils*, *Fibroblasts*, *Epithelial cells*, *T-cells*, *NK-cells*, *HSC-G-CSF*, *Monocyte*, *Macrophage* and *Tissue stem cells*. The *TSCAN* pseudo-time algorithm can be used to define the pseudotemporal pathway of these cell types. Figure 1 illustrates that *TSCAN* identifies a single pseudotemporal pathway that starts with *Tissue stem cells* and ends with *Epithelial cells*.

Step 2. After extracting the pseudotemporal values per cell using the *TSCAN* algorithm, we utilized the *jcglasso* model. To identify key genes for interaction analysis, we focused on those with the highest variability across both tissues (NT and TT) using the Seurat package in R [5]. This resulted in the selection of 260 genes, striking a balance between computational efficiency and the preservation of essential biological information for analysis.

Due to the presence of missing-at-random data, the model used is called *conditional joint missglasso*. This model establishes a unified framework for the two tissue types. The response variable is represented by gene expression and the explanatory variable is represented by pseudo-time. The challenge lies in capturing this nuanced relationship, which does not follow a simple linear path but instead exhibits a piecewise nonlinear pattern. The rationale for this approach is based on the observation that gene expression levels are almost undetectable for pseudo-time values corresponding to cell clusters where genes are under-expressed, but then increase significantly in pseudo-time intervals associated with clusters where they are highly expressed. To account for these dynamics,

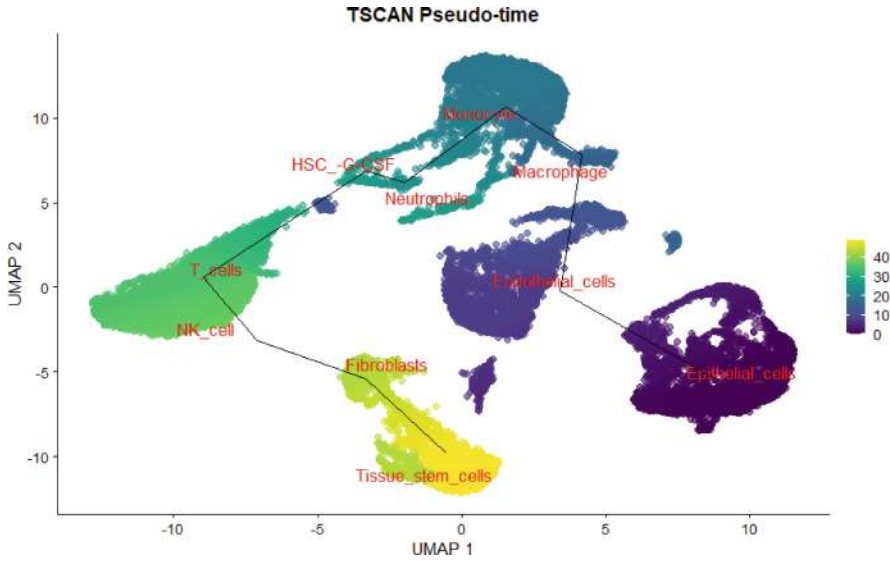


Fig. 1. The graph shows in color scale the pseudotemporal value assigned to each cell by TSCAN. The procedure was carried out and represented in the two-dimensional space generated by UMAP.

we use spline bases to incorporate pseudo-time into the model, with breakpoints set at the beginning and end of the pseudo-time ranges that align with each cell cluster.

Model selection was performed using BIC. Twenty-five values of ρ were used, and the search was then refined around the identified best ρ to enhance accuracy. The α_2 parameter was fixed at 0.5 to equally weigh the influence of both single gene interactions and the dynamics within gene groups in the analysis. Pseudo-time removes genetic interactions that regulate cell differentiation, highlighting only significant disease interactions. The final model obtained a network density of about 6%, and a relevant difference was found between the two examined tissues. In particular, the model estimates a conditional independence between genes for NT, while for TT it highlights the interactions of genes that characterize thyroid cancer (see Fig. 2).

Seven genes have been identified as playing a role in thyroid carcinoma: *UBE2C*, *TOP2A*, *RRM2*, *TK1*, *HMMR*, *H1-5*, and *CDKN3*. Among these genes, *UBE2C* has the most interactions in the obtained network. Its function is focused on ubiquitination, which involves fine regulation of cell cycle and protein stability. Alterations in these processes may contribute to tumorigenesis. The potential utility of *UBE2C* as a target for therapeutic interventions is highlighted by its overexpression in neoplastic tissue [6]. *TOP2A* also plays an important role and has been identified as a central gene in the prognosis of thyroid carcinoma due to its involvement in disease initiation and progression [7]. *RRM2* is a key gene involved in the synthesis of deoxyribonucleotides necessary for DNA replication.

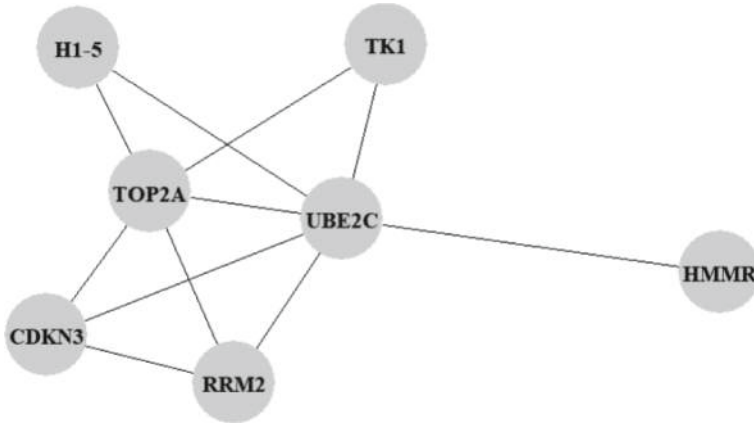


Fig. 2. Igraph representation of the Gene Network in primary thyroid tumor.

Its elevated expression has been linked to adverse outcomes in various types of cancer [8]. *TK1*, which plays a role in nucleotide metabolism and cell proliferation, has an expression profile that correlates with tumor severity. Elevated serum levels of *TK1* in cancer patients indicate disease progression and provide a valuable metric for assessing treatment efficacy and monitoring recurrence [9]. Finally, the research identifies also *HMMR*, *H1-5*, and *CDKN3* as genes significantly associated with thyroid pathology. *HMMR* maintains mitotic spindle integrity and proper chromosomal segregation, *H1-5* is a variant of histone H1 that participates in chromatin compaction and gene expression regulation, and *CDKN3* regulates the cell cycle.

5 Conclusion

This study presents a detailed genetic characterization of thyroid carcinoma in a 17-year-old patient. The genetic profile of the tumor was compared with that of the adjacent healthy thyroid tissue. A *joint conditional missglasso* model was used to analyze genetic interactions across both tissues, with a particular focus on their dependency on pseudo-time, an indicator of the cell differentiation process.

The main innovation of this approach is its ability to isolate interactions between a subset of genes that do not participate in the conventional process of cell differentiation but could still influence tumor pathogenesis. Specifically, the *UBE2C*, *TOP2A*, *RRM2*, and *TK1* genes interact significantly with each other and with other genes, such as *HMMR*, *CDKN3*, and *H1-5*, outlining a distinctive pattern of thyroid carcinoma. In contrast, no interactions were detected between these genes in healthy tissue, indicating complete conditional independence. This highlights the efficiency of pseudo-time in removing any interaction related to the average genetic profile associated with the standard process of cell differentiation.

These results offer new possibilities for future research. By introducing a covariate variable that characterizes the process of cell differentiation, specific disease-associated genetic interactions can be identified. This approach has the potential to significantly impact the study and identification of more precise therapeutic targets and diagnostic biomarkers.

Acknowledgment. Luigi Augugliaro and Gianluca Sottile acknowledge the financial support from Research Projects of National Relevance - PRIN22 (CUP: B53D23009480006).

References

1. Chen, W., et al.: Single-cell RNA-seq reveals MIF-(CD74+ CXCR4) dependent inhibition of macrophages in metastatic papillary thyroid carcinoma. *Oral Oncol.* **148**, 106654 (2024)
2. Ji, Z., Ji, H.: TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**(13), e117–e117 (2016)
3. Sottile, G., Augugliaro, L., Vinciotti, V., Arancio, W., Coronello, C.: Sparse inference of the human hematopoietic system from heterogeneous and partially observed genomic data (2022). arXiv preprint [arXiv:2206.09863](https://arxiv.org/abs/2206.09863)
4. Augugliaro, L., Sottile, G., Wit, E.C., Vinciotti, V.: cglasso: an R package for conditional graphical lasso inference with censored and missing values. *J. Stat. Soft.* **105**, 1–58 (2023)
5. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**(5), 495–502 (2015)
6. Xiang, C., Yan, H.C.: Ubiquitin conjugating enzyme E2 C (UBE2C) may play a dual role involved in the progression of thyroid carcinoma. *Cell Death Discov.* **8**(1), 130 (2022)
7. Gong, M.C., et al.: Prognostic value and significant pathway exploration associated with TOP2A involved in papillary thyroid cancer. *Int. J. Gen. Med.*, 3485–3496 (2021)
8. Qiu, J., et al.: RNA sequencing identifies crucial genes in papillary thyroid carcinoma (PTC) progression. *Exp. Mol. Pathol.* **100**(1), 151–159 (2016)
9. Bitter, E.E., Townsend, M.H., Erickson, R., Allen, C., O'Neill, K.L.: Thymidine kinase 1 through the ages: a comprehensive review. *Cell Biosci.* **10**(1), 1–16 (2020)



Housing Insecurity and First Births in Italy

Alessandro Gallo^(✉) and Daniele Vignoli

University of Florence, Florence, Italy
alessandro.gallo@unifi.it

Abstract. The housing condition is widely recognized as one of the most significant dimensions of individuals' and households' well-being. Consequently, increasing research is being conducted on how to measure housing uncertainty. In this study, we propose a new multidimensional indicator of housing insecurity using a counting approach. We examine the relationship between the proposed indicator and the decision to have the first child. We utilize data from waves spanning 1991 to 2020 from the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy. Preliminary results indicate that housing insecurity is associated with a lower probability of having the first child, particularly among households that do not own a property.

Keywords: Housing Insecurity · Housing Uncertainty · Fertility · Transition to Parenthood

1 Introduction

Housing well-being is receiving increasing attention from scholars and policy-makers. Among the social rights in most OECD countries, the right to adequate housing stands as a fundamental element of well-being [7]. Consequently, indicators of housing uncertainty and deprivation are gaining traction in economic and statistical literature. Moreover, housing shapes social relations, influences access to resources, determines exposure economic segregation [4], and affects intentions and behaviours in the family sphere [10, 11]. We consider the housing insecurity definition as in [4]: the loss of, threat to, or uncertainty of a safe, stable, and affordable home environment.

This work aims to achieve two primary objectives. First, we introduce a novel multidimensional indicator of housing insecurity using a counting approach [1]. This method is commonly employed in diverse fields for generating individual composite indicators and offers several advantages. It simplifies the aggregation of dimensions wherein individuals or households are deemed insecure across various criteria. Subsequently, this method yields a variable indicating overall household or individual insecurity across all dimensions. From this composite individual measure, an aggregated indicator of intensity of housing insecurity

can be easily derived. Additionally, the composite measure allows for decomposition to analyze the distribution of selected dimensions.

Second, we utilize this new measure of housing insecurity to investigate its impact on the probability to have a first child. While extensive research exists on how increasing uncertainty caused the decline of fertility in Italy (but also in the EU) over the past decade [3], less attention has been given to the relationship between fertility and housing insecurity. Among the seminal works linking housing and fertility, [6] found that countries with an unfriendly homeownership regime have the lowest levels of fertility. At the micro level, [11] illustrated a clear positive gradient between the fertility intentions of couples and the degree to which they feel secure about their housing situation. However, a recent work from [10] suggest more complex patterns in the UK. We propose to deepen previous literature by considering housing insecurity as a determinant of the probability to have a first child. Based on existing findings, our hypothesis posits that housing insecurity is related with a reduced probability of having a first child.

We analyze longitudinal data from the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy. We consider a panel spanning from 1991 to 2020 to capture a comprehensive perspective and account for individual unobserved effects in our models. Preliminary results indicate that housing insecurity, when considered in rental situations, is associated with a lower likelihood of transitioning to parenthood.

The remainder of the paper is structured as follows: Sect. 2 outlines the multidimensional indicator using the counting approach, Sect. 3 presents preliminary findings on the relationship between housing insecurity and fertility intentions, and Sect. 4 concludes.

2 A Housing Insecurity Indicator Using a Counting Approach

The counting approach is a methodology proposed by [2] and further developed by [1]. This approach aims to create a set of indicators to measure deprivation or insecurity, and can be applied in various fields. For instance, [9] proposed a multidimensional index to measure economic insecurity in the EU. [5] use this approach to measure employment precariousness using Spanish data from WGS (Wage Structure Survey) or [8] use it to assess well-being using German data from SOEP (Socio-Economic Panel Survey).

Let us consider each individual i belonging to a population N where $N \in \mathbb{N}$. This method works by considering a set of dimensions of insecurity. Once the dimensions are selected, a threshold h_j for each dimension j with $j \in (1, \dots, D)$ is chosen. All the individuals below h_j are insecure in that dimension j . The individual index of Housing Insecurity (H_i) is then given by:

$$0 \leq H_i = \frac{\sum_{j=1}^D w_j I_{ij}}{D} \leq 1 \quad (1)$$

where D is the number of total dimensions j , w_j is the weight given to each dimension j , and I_{ij} is a variable that is equal to 1 if the individual i is insecure in j and 0 if he is not.

Weights w_j can be selected in several ways. One option may be to choose the same weight for each dimension. A second option following [9] could be to assign a weight equal to the proportion of individuals not suffering from that dimension of insecurity. In this case, weights are given by:

$$w_j = \frac{D \cdot P_j}{\sum_{j=1}^D P_j} \quad (2)$$

where P_j is the portion of population not suffering of the dimension j insecurity.

The following step is to choose a *general threshold* k to identify which individuals are housing insecure. Here we select the number of dimensions sufficient to identify an individual as housing insecure. There are several strategies to identify k . This needs to be chosen according to the context.

2.1 Aggregate Index

One of the advantages of this approach is that the individual index can be easily grouped to compute several aggregated indices. In this work we present an index of prevalence of housing insecurity (PHI) and an index of intensity of housing insecurity ($\mu_{HI}^{q_{HI}}$).

The prevalence of housing insecurity (PHI) is given by the share of insecure in the population. It can be computed as follows:

$$PHI = \frac{\sum_{i=1}^N I(H_i \geq k)}{N} = \frac{q_{HI}}{N}. \quad (3)$$

$I(H_i \geq k)$ takes the value 1 if an individual is housing insecure while is 0 if the individual is not. q_{HI} is the number of individuals insecure in the population N . The intensity of housing insecurity ($\mu_{HI}^{q_{HI}}$) the intensity of housing insecurity among the insecure individuals. It can be measured by the mean value of insecurity among the insecure and is given by:

$$\mu_{HI}^{q_{HI}} = \frac{\sum_{i=1}^N H_i I(H_i \geq k)}{\sum_{i=1}^N I(H_i \geq k)} = \frac{\sum_{i=1}^N H_i I(H_i \geq k)}{q_{HI}}. \quad (4)$$

2.2 The Dimensions of Housing Insecurity

Tenants and *Homeowners* may have different needs according to the type of housing condition in which they are in. For this reason, we propose two indicators, one for Tenants' share of the population (HIT) and one for the Homeowners' share (HIO). We selected $D = 5$ dimensions for each indicator. All the dimensions have a specific threshold h_j chosen according to the context. There is a set of 3 dimensions that are common among both indicators and then a set of 2 dimensions specific for each indicator.

The common dimensions are: i) If the amount of purchases of other durable goods (furniture, furnishings, appliances, and so forth) is larger than 30% of Household Disposable Income. ii) The presence during the year of moments or periods in which the family has been in arrears for an extended period (at least 90 days) with the payment of bills (gas, electricity, telephone, etc.). iii) If the location (area) of the residence is in a particularly degraded area.

The specific dimensions for the *Tenants* are: i) If the annual rent paid is larger than 30% of Household Disposable Income. ii) The presence during the year of moments or periods in which the family has been in arrears for an extended period (at least 90 days) with the payment of rent for the dwelling in which they live. In the case of the *Homeowners*: i) If the total amount of installments paid during the year for the purchase or renovation of owned real estate is larger than 30% of annual household disposable income. ii) If the amount of extraordinary maintenance expenses on owned properties is larger than the 30% of Household Disposable Income.

The dimensions of *HIT* and *HIO* have weights w_j according to Eq. (2). This introduces the concept of relativity, fostering comparison dynamics within society. The greater the proportion of individuals unaffected by a dimension, the greater the perceived insecurity for those who do experience it. An exception is the first common dimension regarding the durable goods. Since we evaluate that its impact on housing insecurity is lower, we consider a weight of 0.5. The general threshold k is 0.2 for *HIT* and 0.21 for *HIO*. With this values to be considered housing insecure, individuals need to have at least two dimensions of insecurity. Alternatively they may have just one among the specific dimensions or they need to live in a particularly degraded area.

To measure housing insecurity on a broader perspective a proposal may be to consider an indicator based on a union between the two indices. This indicator would be a dummy variable equal to 1 if *HIT* or *HIO* are equal to one and 0 if both are 0. Formally the general indicator *HI* is given by:

$$HI_i = \begin{cases} 1 & \text{if } HIO_i = 1 \text{ or } HIT_i = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

3 Effects of Housing Insecurity on First Births in Italy

To study the relation between housing insecurity and fertility in Italy, we consider data from the Survey on Household Income and Wealth (SHIW) carried out by the Bank of Italy. We use a time span ranging from 1991 to 2020. Our statistical unit are heterosexual couples with no children, living on the same home. Their age is included between 18 and 54 years for the male partner and between 18 and 45 years for the female one. They also need to remain in the survey for at least two waves since all the covariates of the model are lagged to the previous wave to be coherent with the conception of a child. Our estimation sample is composed by 1028 couple/years. The dependent variable is a dummy equal to 1 if they have a child and 0 if not.

The main independent variable is the *HI* computed as described in the previous section. The controls we use are the standard socio-demographic variables used in the literature, considered for both partners. We include age (also in squared terms to control for non-linearity), education (measured as a dummy that states 1 if they have a degree) and the labor market status (a categorical variable equal to 1 if they are employees, 2 if they are self-employed and 3 if they are not employed). We also include a dummy variable to indicate if the couple is renting their home or they own it, to differentiate *Tenants* and *Homeowners*. We consider year fixed-effects and regional fixed-effects and finally individual unobserved random effects to control for unobserved personality traits of the couples.

Table 1. Preliminary Results: Logit coefficients estimates on first births.

	(1)	(2)	(3)
Housing Insecurity (HI_i)	0.585 (0.365)	0.469 (0.369)	0.460 (0.423)
Renting	0.212 (0.163)	0.026 (0.169)	0.027 (0.177)
Housing Insecurity*Renting	-1.299** (0.540)	-1.098* (0.573)	-1.134* (0.611)
<i>Individual Controls</i>	-	Yes	Yes
<i>Individual Random Effects</i>	-	-	Yes
<i>N</i>	1028	1028	1028

Notes: Standard errors in parentheses are clustered at the household level. The dependent variable is a dummy equal to one if the couple has the first child. Individual controls are age, educational level and employment status.

3.1 Preliminary Results

Coefficient estimates are shown in Table 1. In model (1) the covariates are: the *HI*, the dummy describing the housing status, the interaction terms with *HI* to asses the differences in the effect among *Homeowners* and *Tenants*, and the wave effects and the regional effects. In the second model individual controls are added, and finally in the third column the full model with the individual unobserved random effects is presented. These preliminary results suggest that there is a significant negative effect of Housing insecurity on the probability of having the first child in the case of *Tenants*. This effect remain significant also in the second model and in the third one, when further degrees of complexity are added. To gain a substantial interpretation of the effect sizes estimated, marginal effects on the probability to have the first child are computed. For instance, being

housing insecure when the couples are *Tenants* produce a 10% lower probability of having the first child.

These preliminary results confirm our hypothesis that housing insecurity is associated to a lower probability of becoming parents.

4 Conclusions

In this paper we propose a new indicator of housing insecurity using a counting approach. This approach considers a set of dimensions of insecurity and counts the number of insecurity dimensions in which each individual/household is insecure. We consider this composite measure in the context of fertility behaviours in Italy. Our hypothesis is that being housing insecure is associated with a lower probability of having the first child. The estimated results show that for *Tenants* couples, if they are housing insecure, the probability of becoming parents is 10% lower then for the not insecure.

Further developments may be done by considering also the transition to higher-order births.

References

1. Alkire, S., Foster, J.: Counting and multidimensional poverty measurement. *J. Public Econ.* **95**(7–8), 476–487 (2011). <https://doi.org/10.1016/j.jpubeco.2010.11.00>
2. Atkinson, A.B.: Multidimensional insecurity: contrasting social welfare and counting approaches. *J. Econ. Inequality* **1**, 51–65 (2003). <https://doi.org/10.1023/A:1023903525276>
3. Comolli, C.L., Vignoli, D.: Spreading uncertainty, shrinking birth rates: a natural experiment for Italy. *Eur. Sociol. Rev.* **37**(4), 555–570 (2021). <https://doi.org/10.1093/esr/jcab001>
4. De Luca, S., Rosen, E.: Housing insecurity among the poor today. *Ann. Rev. Sociol.* **48**, 343–371 (2022). <https://doi.org/10.1146/annurev-soc-090921-040646>
5. García-Pérez, C., Prieto-Alaiz, M., Simón, H.: A new multidimensional approach to measuring precarious employment. *Soc. Ind. Res.* **134**, 437–454 (2017). <https://doi.org/10.1007/s11205-016-1447-6>
6. Mulder, C.H., Billari, F.C.: Homeownership regimes and low fertility. *Hous. Stud.* **25**(4), 527–541 (2010). <https://doi.org/10.1080/02673031003711469>
7. Navarro Ruiz, C., Ayala, L.: Multidimensional housing deprivation indices with application to Spain. *Appl. Econ.* **40**(5), 597–611 (2008). <https://doi.org/10.1080/00036840600722323>
8. Peichl, A., Pestel, N.: Multidimensional well-being at the top: evidence for Germany. *Fiscal Stud.* **34**(3), 355–371 (2013). <https://doi.org/10.1111/j.1475-5890.2013.12010.x>
9. Romaguera de la Cruz, M.: Measuring economic insecurity using a counting approach. An application to three european countries. *Rev. Income Wealth* **66**(3), 558–583 (2020). <https://doi.org/10.1111/roiw.12428>

10. Tocchioni, V., Berrington, A., Vignoli, D., Vitali, A.: The changing association between homeownership and the transition to parenthood. *Demography* **58**(5), 1843–1865 (2021). <https://doi.org/10.1215/00703370-9420322>
11. Vignoli, D., Rinesi, F., Mussino, E.: A home to plan the first child? Fertility intentions and housing conditions in Italy. *Popul. Place Space* **19**, 60–71 (2013). <https://doi.org/10.1002/psp.1716>



Investigating Auxiliary Mixture Sampling for Poisson Regression Models

Aldo Gardini^(✉), Fedele Greco, and Carlo Trivisano

Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italy
`aldo.gardini@unibo.it`

Abstract. In this paper we revise a popular alternative for estimating Poisson regression models in a Bayesian framework and discuss possible pitfalls tied to data features. The MCMC algorithms are based on augmenting the model via the introduction of auxiliary variables. This leads to a model linear in the regression parameters with errors following a Gumbel or a log-Gamma distribution depending on the augmentation strategy. Such distributions are approximated by a Gaussian Mixture in order to favor standard MCMC sampling with Gibbs steps after augmentation. We show situations when such an approximation deteriorates and causes non-convergence of the algorithm, discussing how this can be detected while the algorithm is running.

Keywords: Bayesian inference · Gibbs samplers · mixture of Gaussian · disease mapping

1 Introduction

The problem of estimating Poisson regression models in the Bayesian paradigm has motivated several research works in the field of computational statistics, to propose efficient Monte Carlo Markov Chain (MCMC) algorithms. Indeed, due to the lack of conditional conjugacy under any prior setting, the use of Metropolis-Hastings or Hamiltonian Monte Carlo algorithms is required, and some strategies for the Poisson model have been proposed for particular contexts (e.g. spatial models [7]).

Alternatively, a popular strand of literature focuses on the adoption of data augmentation schemes, followed by an approximation of the model residuals posterior distributions using a mixture of Gaussian distributions [4, 5]. In this framework, it is worth mentioning [3], which adopted an alternative approximation strategy.

Here we focus on the data augmentation schemes proposed in [4, 5]. Such strategies are widely used as they allow to resort to Gibbs samplers since the full conditionals of the coefficients and random effects in Latent Gaussian Models reduce to Gaussian distributions. As a consequence, the algorithms are easy to implement and, for instance, many R packages exploit them to fit Poisson models [2, 8]. Despite an analysis of the efficiency of such algorithms in a spatial

regression framework was carried out in [6], our goal is to show that a careful check of the approximation validity is required, as lack of convergence can be observed in some situations where approximation's accuracy deteriorates.

2 Basic Concepts on the Sampling Algorithms

Consider a basic Poisson regression model specified for a sample of counts:

$$y_i | \lambda_i \stackrel{ind}{\sim} \mathcal{P}(\lambda_i), \quad i = 1, \dots, n; \quad (1)$$

assuming

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}. \quad (2)$$

Note that, the basic ideas on the sampler do not change if both random effects and fixed effects are included in the linear predictor. We assume that the prior on the coefficients vector is $\boldsymbol{\beta} \sim \mathcal{N}_P(\mathbf{0}, \mathbf{V}_0)$. Exploiting the properties of the Poisson process, it is possible to express such a model as the sequence of inter-arrival times, leading to two possible data augmentation schemes discussed in what follows.

2.1 Basic Data Augmentation Scheme

To express the Poisson distribution in terms of inter-arrival times, it is required to introduce $y_i + 1$ independent Exponential random variables with intensity parameter λ_i for each sampling count y_i , $i = 1, \dots, n$. This sequence of random variables is defined as

$$\tau_{ij} | \lambda_i = \frac{\zeta_{ij}}{\lambda_i}; \quad j = 1, \dots, y_i + 1; \quad i = 1, \dots, n;$$

where $\zeta_{ij} \stackrel{ind}{\sim} \text{Exp}(1)$, $\forall i, j$. Taking the negative of the logarithm, the model can be expressed through the following linear relationship involving the coefficients:

$$y_{ij}^* | \boldsymbol{\beta} = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_{ij}, \quad (3)$$

where $y_{ij}^* = -\log(\tau_{ij})$ is the auxiliary variable determining the linear model and $\varepsilon_{ij} = -\log(\zeta_{ij}) \sim \text{Gumbel}(0, 1)$. In this way, $n^* = n + \sum_{i=1}^n y_i$ auxiliary variables are introduced. [4] proposed to introduce a second step of data augmentation by approximating the density function of the Gumbel random variable (or Type I Extreme Value distribution)

$$f_{\varepsilon_{ij}}(z) = \exp \{ -z - e^{-z} \}, \quad \forall i, j,$$

through a mixture of Gaussian distributions

$$f_{\varepsilon_{ij}}(z) = \sum_{r=1}^k w_r \phi(z; m_r, s_r^2), \quad (4)$$

i.e. including in the model a new error term $\tilde{\varepsilon}_{ij}$. The simple normal linear model emerges by expressing the mixture conditionally on the labels $r_{ij} \in \{1, \dots, k\}$:

$$y_{ij}|\beta, r_{ij} = \mathbf{x}_i^\top \beta + m_{r_{ij}} + \tilde{\varepsilon}_{ij}, \quad \tilde{\varepsilon}_{ij}|r_{ij} \sim \mathcal{N}(0, s_{r_{ij}}^2).$$

To express the model in matrix form, we need to set some further notation. The whole set of labels is defined as $\mathbf{r} \in \{1, \dots, k\}^{n^*}$, consequently we define also the sets of means and variances of the mixture components conditionally on the specific labels $\mathbf{m}(\mathbf{r}) = (m_{r_{11}}, \dots, m_{r_{n, y_n}})$ and $\mathbf{s}^2(\mathbf{r}) = (s_{r_{11}}^2, \dots, s_{r_{n, y_n}}^2)$. Storing the auxiliary variables into $\mathbf{y}^*$ and the related auxiliary information in $\mathbf{X}^* \in \mathbb{R}^{n^* \times P}$, we can express the linear model conditional on the latent labels indicating the mixture components:

$$\mathbf{y}^* = \mathbf{X}^* \beta + \mathbf{m}(\mathbf{r}) + \tilde{\varepsilon}, \quad \tilde{\varepsilon}|\mathbf{r} \sim \mathcal{N}_{n^*}(\mathbf{0}, \mathbf{D}(\mathbf{r})),$$

where $\mathbf{D}(\mathbf{r}) = \text{diag}(\mathbf{s}^2(\mathbf{r}))$.

The MCMC algorithm is implemented by iterating thorough the following steps:

- *Step 1.* Draw β from the full conditional, under the approximated model $f_{\beta|\mathbf{r}, \tau}$:

$$\beta|\mathbf{r}, \mathbf{y}^* \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \quad (5)$$

where $\boldsymbol{\Sigma}_\beta = (\mathbf{X}^{*\top} \mathbf{D}(\mathbf{r})^{-1} \mathbf{X}^* + \mathbf{V}_0^{-1})^{-1}$ and $\boldsymbol{\mu}_\beta = \boldsymbol{\Sigma}_\beta \mathbf{X}^{*\top} \mathbf{D}(\mathbf{r})^{-1} (\mathbf{y}^* - \mathbf{m}(\mathbf{r}))$.

- *Step 2.* To draw the n^* auxiliary variables $\mathbf{y}^*$, the fact that the arrival times of a Poisson process conditioned on having observed a given number of jumps are distributed as the order statistics of a sample from a Uniform distribution on the interval $[0, 1]$ is exploited. In addition, the $(y_i + 1)$ -th arrival time after the unit time is drawn exploiting that $\tau_{i(y_i+1)} = 1 - \sum_{j=1}^{y_i} \tau_{ij} + \zeta_i/\lambda_i$, where $\zeta_i \sim \text{Exp}(1)$.
- *Step 3.* The component indices r_{ij} are sampled with probabilities proportional to $\mathbb{P}[r_{ij} = k | y_{ij}^*, \beta]$.

2.2 Improved Data Augmentation Scheme

The main drawback of the previously described data augmentation scheme is the large number of latent variables that are introduced, that is $\sum_{i=1}^n y_i + n$, particularly when the overall number of observed counts is huge. To address this issue, [5] proposed an alternative and more parsimonious scheme. For each observation, the latent variable τ_{1i}^* is considered to play the role of the $(y_i + 1)$ -th inter-arrival time after time 1, whereas the variable τ_{2i}^* has the role of the y_i -th arrival time. Note that, when $y_i = 0$ such an arrival time is 0, and no latent variable is required. While the set of latent variables τ_{1i}^* is basically equal to those presented in the previous section, the distribution of the second latent variable follows from the fact that the arrival time is the sum of y_i independent Exponential distributions with rate λ_i , hence:

$$\tau_{2i}^* = \frac{\zeta_{2i}^*}{\lambda_i}, \quad \zeta_{2i}^* \sim \text{Gamma}(y_i, 1).$$

Hence, a new mixture of Gaussians approximation to this distribution was introduced in [5].

3 Some Examples

3.1 A Toy Example

In this Section, a toy example is discussed, to show how the issues affecting the approximation behind the auxiliary mixture sampling algorithms can emerge. To this aim, we generated a dataset with $n = 30$ observations under a Poisson model, considering two distinct covariates. Then, only the first covariate is included in the regression model, yielding a misspecified model.

As a benchmark, we fit the Poisson regression model with INLA and Stan, comparing the resulting posterior distribution of the regression coefficient β_1 with those obtained under the auxiliary mixture sampling algorithms (Fig. 1).

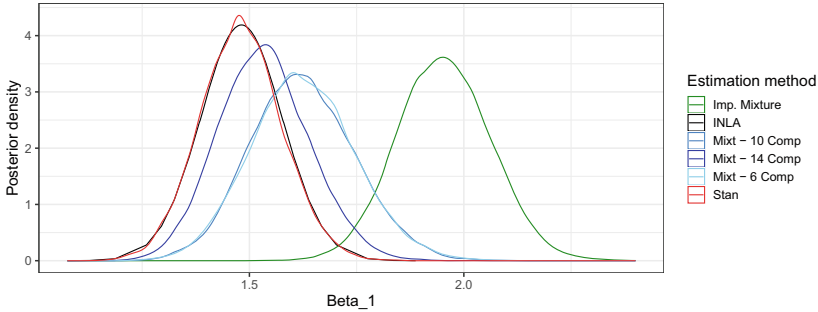


Fig. 1. Posterior distribution of β_1 obtained under the benchmark methods (INLA and Stan), under the auxiliary sampling scheme of [4] with $k \in \{6, 10, 14\}$ and the improved sampling scheme of [5].

As a first outcome, it is possible to notice that the approximation adopted by the improved sampling scheme [5] leads to a posterior distribution notably far from the ones determined by INLA and Stan. This points out that, even if the improved scheme is more computationally efficient, due to the reduced number of auxiliary variables, it seems to be more prone to errors when approximating the distributions of the residuals. Conversely, the posteriors under the original sampling scheme by [4] are closer to the benchmark posteriors. In this case, we also tested different numbers of components k of mixture approximation (4), comparing the results obtained with $k = 10$, which is advised by the authors, to the results obtained with $k = 6$ and $k = 14$. From this exercise, we can notice that the accuracy of the approximation can matter when some auxiliary residuals are in the tails of the distribution. Indeed, when the approximation is improved ($k = 14$), the posterior of β_1 approaches the benchmark.

3.2 A Real Data Example: the BYM Model

In this Section, we apply the Besag York and Mollié (BYM) model [1] to death counts for Diabetes and Tumors in the $n = 341$ municipalities of the Emilia Romagna Region between 2004 and 2009. Let y_i and E_i denote observed and expected counts respectively. The model is specified as

$$y_i | E_i, \lambda_i \sim \text{Poisson}(E_i \lambda_i) \quad i = 1, \dots, n$$

$$\log(\lambda_i) = \alpha + u_i + v_i$$

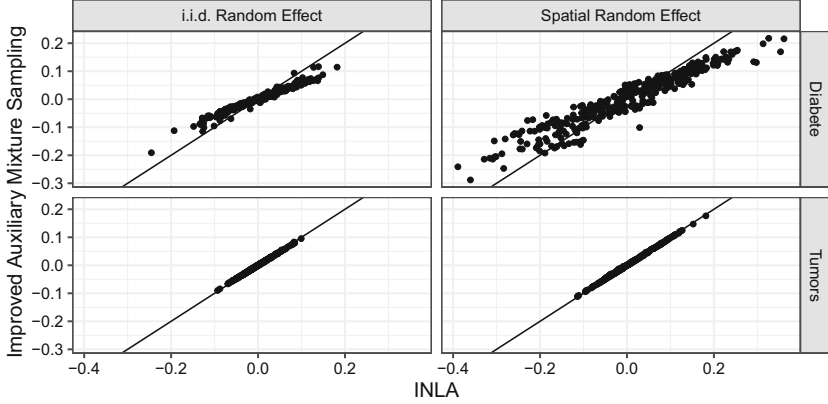


Fig. 2. Posterior means of unstructured (left) and structured (right) random effects for Diabetes (upper panels) and Tumors (bottom panels). Auxiliary Mixture Sampling vs INLA.

where $\mathbf{u} = (u_1, \dots, u_i, \dots, u_n)^\top \sim \mathcal{N}_n(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$ is a vector of unstructured (*iid*) random effects, while $\mathbf{v} = (v_1, \dots, v_i, \dots, v_n)^\top \sim \text{ICAR}(\sigma_v^{-2} \mathbf{K}_v)$. Vector $\mathbf{v}$ denotes spatially structured random effects following an Intrinsic Conditional Autoregressive distribution whose rank-deficient precision matrix is specified as $\mathbf{K}_v = \mathbf{D} - \mathbf{W}$ where $w_{ij} = 1$ if municipalities i and j share a common boundary while $w_{ij} = 0$ otherwise and $\mathbf{D}$ is a diagonal matrix with diagonal entries corresponding to the number of neighbors of the i -th municipality. Both random effects are subject to a sum-to-zero constraint for the sake of identifiability.

In Fig. 2, posterior means for both structured and unstructured random effects obtained with INLA and improved auxiliary mixture sampling are compared. Perfect agreement between the two estimation methods can be noticed in the case of Tumors data (showing a total of 70274 observed deaths). On the other hand, when considering Diabetes deaths (4877 observed deaths in the study period), it can be seen that posterior means obtained with the two methods are in disagreement, again highlighting some difficulties of the Improved algorithm to converge to the correct posterior distribution when data contain some observation leading to extreme auxiliary residuals with respect to mixture approximation (4).

4 Concluding Remarks

The examples discussed in the previous section allow to point out that particular care is required when a Poisson model is fit by means of the auxiliary sampling schemes introduced in [4] and [5]. Indeed, the mixture approximations of the auxiliary residuals distribution guarantee the convergence to the correct stationary distribution only when the model assumptions are fulfilled by the data. For this reason, we advise monitoring indicators that assess goodness of the approximation while running the algorithm (e.g., based on the comparison of the likelihood of the approximated and true models), and, if required, adding a rejection step to foster convergence to the target posterior distribution.

References

1. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**(1), 1–20 (1991)
2. Dvorzak, M., Wagner, H.: Sparse bayesian modelling of underreported count data. *Stat. Model.* **16**(1), 24–46 (2016)
3. D’Angelo, L., Canale, A.: Efficient posterior sampling for bayesian poisson regression. *J. Comput. Graph. Stat.* **32**(3), 917–926 (2023)
4. Frühwirth-Schnatter, S., Wagner, H.: Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika* **93**(4), 827–841 (2006)
5. Frühwirth-Schnatter, S., Frühwirth, R., Held, L., Rue, H.: Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Stat. Comput.* **19**, 479–492 (2009)
6. Gschlößl, S., Czado, C.: Does a gibbs sampler approach to spatial poisson regression models outperform a single site MH sampler? *Comput. Stat. Data Anal.* **52**(9), 4184–4202 (2008)
7. Knorr-Held, L., Rue, H.: On block updating in markov random field models for disease mapping. *Scand. J. Stat.* **29**(4), 597–614 (2002)
8. Martin, A.D., Quinn, K.M., Park, J.H.: Mcmcpack: Markov chain Monte Carlo in R. *J. Stat. Softw.* **42**(9), 1–21 (2011)



Unobserved Heterogeneity and Identification of Causal Effects Using Mixture Models

Leonardo Genesin^{1(✉)}, Marco Bertoni², Adriano Paggiaro¹,
and Giovanna Menardi¹

¹ Department of Statistical Sciences, University of Padova, Padua, Italy
leonardo.genesin@phd.unipd.it

² Department of Economics and Management “Marco Fanno”, University of Padova,
Padua, Italy

Abstract. We explore the opportunity of identifying potential sources of latent heterogeneity by employing a finite mixture regression model aimed at evaluating the causal effect of a treatment. Within each latent class of heterogeneity, the model describes the relationship between the outcome and the treatment, controlling for a set of covariates, and manages the endogeneity of the treatment. The proposed methodology can be applied to data of arbitrary nature, also allowing for the identification of possible heterogeneous effects.

Keywords: causal inference · latent variables · mixture model

1 Introduction

Regression models traditionally represent a simple framework to formalize the problem of evaluating the causal effect of a treatment on some outcome of interest. For the sake of simplicity, consider the setting defined by a linear model

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \quad (1)$$

where $Y_i \in \mathbb{R}$ is the outcome, observed on subject i , $i = 1, \dots, n$, and $D_i \in \{0, 1\}$ is the exposure variable. The opportunity of interpreting β_1 as the average treatment effect (ATE) relies on the exogeneity of D , i.e. when D is assigned randomly to the subjects, as in randomized control trials. In fact, in observational studies, the outcome often depends on some factors which also affect the probability of exposure. Omitting such factors from (1) entails D to be correlated with the error term ε , and the Ordinary Least Squares (OLS) estimator of β_1 to be biased, hence precluding its interpretation in terms of ATE [1]. Nowadays, the ever-increasing availability of data and variables allows to control for many confounding factors so that, conditionally on observed characteristics, it is not unplausible to assume the treatment assignment to be independent of the error term. In fact, there still often exists a residual source of unobservable heterogeneity.

Motivated precisely in contexts where the variable of interest exhibits some form of heterogeneity within the population, mixture models represent a class

of flexible statistical models used to describe data arising from distinct, unobservable groups, often referred to as latent, and each characterized by its own probability distribution [2].

In the light of this link, the aim of this work is to explore the opportunity of using mixture models in a context of causal inference, to identify possible sources of heterogeneity and aid the evaluation of the causal effect of a treatment on some outcome of interest. After presenting the model and its main assumptions, in the following we draft an Expectation Maximization (EM) algorithm specifically designed for determining the maximum likelihood estimate of the parameters, and we illustrate its application to a set of data collected via the Survey of Health, Ageing and Retirement in Europe (SHARE).

2 The Proposed Model

2.1 Model Assumptions and Specification

We build on the idea that, even after accounting for any observed attributes, there exist groups within the population, associated with some unobserved characteristic, and within such groups the treatment can be regarded as randomly assigned. Formally, in the linear model formulation described by Eq. (1), we denote by (X, Z) a vector of control variables, assumed to describe all the heterogeneity of the population for different treatment assignments. Consistently with the usual framework of causal inference models, we assume that, for a subject i , $D_i \perp \varepsilon_i | (X_i, Z_i)$. In fact, unlike the usual Conditional Independence Assumption (CIA) [1], while $X = (X_1, \dots, X_p)^T$ is a vector of known, observed control variables, Z is unobservable and thereby latent.

Since Z describes the residual heterogeneity, arising from possible different sources, it makes sense to define it by means of an unordinal, categorical variable, whose levels catch possible confounding factors having different nature and influence on both the treatment and the outcome. Hence, we let $Z_i = (Z_{i1}, \dots, Z_{iG})$ with $Z_{ig} = 1$ when subject i belongs to class g , $g = 1, \dots, G$, 0 otherwise.

The proposed methodology is defined by three simultaneous equations which model, conditionally to the class membership, a possibly different relationship between Y and the treatment D , the probability of exposure, and the probability of the latent class, given the control variables. The methodology, easily extendable to alternative specifications, is here illustrated with reference to a linear model, i.e.

$$Y|(Z_g = 1) = \beta_{0g} + \beta_{1g}D + X^T\beta_{2g} + \varepsilon_g, \quad g = 1, \dots, G.$$

Hence, assuming $\mathbb{E}(\varepsilon_g) = 0$, $\mathbb{V}(\varepsilon_g) = \sigma_g^2$, and $\mu_g = \mu_g(X) = \mathbb{E}(Y|Z_g = 1, D, X)$,

$$\begin{cases} \mu_g &= \beta_{0g} + \beta_{1g}D + X^T\beta_{2g} \\ \log\left(\frac{\rho_g}{1-\rho_g}\right) &= \delta_{0g} + X^T\delta_{1g} & g = 1, \dots, G \\ \log\left(\frac{\pi_g}{\pi_1}\right) &= \gamma_{0g} + X^T\gamma_{1g}, & g = 2, \dots, G. \end{cases} \quad (2)$$

Estimation via maximum likelihood requires a suitable specification of the following underlying distributions:

$$\begin{cases} Y|Z_g = 1, D = d, X = x \sim p_g(y|d, x, \theta_g), & \theta_g = (\beta_{0g}, \beta_{1g}, \beta_{2g}, \sigma_g^2) \\ D|Z_g = 1, X = x \sim p_D(d|x, \delta_g) & \delta_g = (\delta_{0g}, \delta_{1g}) \\ Z|X = x \sim p_Z(z|x, \gamma) & \gamma = (\gamma_{02}, \gamma_{12}, \dots, \gamma_{0G}, \gamma_{1G}). \end{cases}$$

Given Eq. (2), the most natural choices turn out to be, respectively, $Y|Z_g = 1, D, X \sim N(\mu_g, \sigma_g^2) = N(\beta_{0g} + \beta_{1g}D + X^T \beta_{2g}, \sigma_g^2)$; $D|Z_g = 1, X \sim Bi(1, \rho_g) = Bi\left(1, \frac{e^{\gamma_{0g} + X^T \gamma_{1g}}}{1 + e^{\gamma_{0g} + X^T \gamma_{1g}}}\right)$, and $Z|X \sim Bi_G(1, \pi)$, with $\pi = (\pi_1, \dots, \pi_G)$, $\pi_1 = \frac{1}{1 + \sum_{h=2}^G \frac{e^{\gamma_{0h} + X^T \gamma_{1h}}}{e^{\gamma_{0h} + X^T \gamma_{1h}}}}$ and $\pi_g = \frac{e^{\gamma_{0g} + X^T \gamma_{1g}}}{1 + \sum_{h=2}^G \frac{e^{\gamma_{0h} + X^T \gamma_{1h}}}{e^{\gamma_{0h} + X^T \gamma_{1h}}}}$, $g = 2, \dots, G$.

2.2 Model Estimation

Given a sample $(y, d, x) = \{(y_i, d_i, x_i)\}_{i=1, \dots, n}$, the parameters of (2) are estimated via maximum likelihood, by means of a suitably designed EM algorithm. To this end, the expectation of the complete likelihood over the latent groups is

$$\begin{aligned} Q(\gamma, \delta, \theta) &= \mathbb{E}_Z [\ell^C(\gamma, \delta, \theta|y, d, x)] \\ &= \sum_{i=1}^n \sum_{g=1}^G \mathbb{E}_Z [Z_{ig} \log [\pi_g p_D(d_i|x_i, \delta_g) p_g(y_i|d_i, x_i, \theta_g)]] , \\ &= \sum_{i=1}^n \sum_{g=1}^G \mathbb{E}_Z [Z_{ig}|y, d, x] \log [\pi_g p_D(d_i|x_i, \delta_g) p_g(y_i|d_i, x_i, \theta_g)] , \\ &= \sum_{i=1}^n \sum_{g=1}^G \tau_{ig} [\log \pi_g + \log p_D(d_i|x_i, \delta_g) + \log p_g(y_i|d_i, x_i, \theta_g)] . \quad (3) \end{aligned}$$

This is maximized by alternating an E-step and an M- step until convergence. At the k -th iteration:

E-step. Given the observed data, $(\gamma^{(k-1)}, \delta^{(k-1)}, \theta^{(k-1)})$, and $\pi_g^{(k-1)}, \rho_g^{(k-1)}$, we compute expectation in (3) by:

$$\begin{aligned} \tau_{ig}^{(k)} &= \mathbb{E} [Z_{ig}|y, d, x] = P(Z_{ig} = 1|y, d, x) \\ &= \frac{p(Z_{ig} = 1|x_i) p_D(d_i|Z_{ig} = 1, x_i, \delta_g^{(k-1)}) p_g(y_i|Z_{ig} = 1, d_i, x_i, \theta_g^{(k-1)})}{\sum_{g=1}^G p(Z_{ig} = 1|x_i) p_D(d_i|Z_{ig} = 1, x_i, \delta_g^{(k-1)}) p_g(y_i|Z_{ig} = 1, d_i, x_i, \theta_g^{(k-1)})} \\ &= \frac{\pi_g^{(k-1)} \rho_g^{(k-1)} p_g(y_i|Z_{ig} = 1, d_i, x_i, \theta_g^{(k-1)})}{\sum_{g=1}^G \pi_g^{(k-1)} \rho_g^{(k-1)} p_g(y_i|Z_{ig} = 1, d_i, x_i, \theta_g^{(k-1)})}, \quad g = 1, \dots, G, \quad i = 1, \dots, n. \end{aligned}$$

Note that the presence of the probabilities ρ_g in the computation of the posterior probabilities τ_{ig} highlights how the specification of the second equation

in (2) to describe the treatment assignment conditionally on control variables and Z aids in the identification of latent heterogeneity.

M-step. Given the $\tau_{ig}^{(k)}$, computed in the E-step, we update the parameters by maximizing the (3).

$$\begin{aligned} (\gamma^{(k)}, \delta^{(k)}, \theta^{(k)}) &= \operatorname{argmax}_{(\gamma, \delta, \theta)} Q(\gamma, \delta, \theta; \tau_{ig}^{(k)}) \\ &= \operatorname{argmax}_{(\gamma, \delta, \theta)} \sum_{i=1}^n \sum_{g=1}^G \tau_{ig}^{(k)} [\log \pi_g + \log p_D(d_i | x_i, \delta_g) + \log p_g(y_i | d_i, x_i, \theta_g)] \end{aligned} \quad (4)$$

It is worth noting that, despite the increased complexity of the model, which involves, with respect to a simple regression mixture, the inclusion of an additional equation to explain the relationship between D and the control variables within each group, the function $Q(\cdot)$ is still separable in parameters, and

$$\begin{aligned} \theta^{(k)} &= \operatorname{argmax}_{\theta} \sum_{i=1}^n \sum_{g=1}^G \tau_{ig}^{(k)} \log p_g(y_i | d_i, x_i, \theta_g) \\ \delta^{(k)} &= \operatorname{argmax}_{\delta} \sum_{i=1}^n \sum_{g=1}^G \tau_{ig}^{(k)} \log p_D(d_i | x_i, \delta_g) \\ \gamma^{(k)} &= \operatorname{argmax}_{\gamma} \sum_{i=1}^n \sum_{g=1}^G \tau_{ig}^{(k)} \log \pi_g \end{aligned}$$

whose solution depends on the specific choices for p_g , p_D , and p_Z .

In the specific case of Normal linear model, it can be easily shown that the update mechanism of $\theta_g = (\mu_g, \sigma_g^2) = (\beta_{0g}, \beta_{1g}, \beta_{2g}, \sigma_g^2)$ boils down to an explicit solution via weighted least squares [3], hence

$$\beta_g^{(k)} = (X^T W_g^{(k)} X)^{-1} X^T W_g^{(k)} y, \quad g = 1, \dots, G$$

with $(\beta_{0g}, \beta_{1g}, \beta_{2g})$, and $W_g^{(k)}$ a diagonal matrix with elements $\tau_{ig}^{(k)}$ for $i = 1, \dots, n$; furthermore, the variances σ_g^2 are updated by:

$$\sigma_g^{2(k)} = \frac{\sum_{i=1}^n \tau_{ig}^{(k)} (y_i - X \beta_g^{(k)})^2}{\sum_{i=1}^n \tau_{ig}^{(k)}}, \quad g = 1, \dots, G.$$

The parameters δ_g , involved in modelling the treatment assignment versus the observed controls, require an update step whose computational demand depends on the distribution assumed for D . In the specific case of binary treatment, or more generally when D belongs to the exponential family, the estimating equation do not generally admit explicit solution. However, apart from minor adjustments due to the presence of the weights τ_{ig} , the maximization can be reduced to using the iterated weighted least squares (IWLS) algorithm, typical of generalized linear models.

Estimation of the γ parameters linking the control variables to the mixture weights π_g does not generally admit closed form solutions and requires estimation algorithms usually employed in multinomial regression models. See, e.g., [4].

After estimating effects conditional on X and Z , ATE is computed by averaging so that groups that have more treated get higher weight.

3 A Real Data Example

We present the results of the application of the proposed model on a set of data collected via SHARE¹, a survey aimed at exploring several socio-economic aspects of a number of european citizens older than 50. The goal of the specific application is to estimate the effect of Education (in years) on the earnings (measured by the current wage). We benchmark results obtained via OLS and Two-Stage Least Squares (2SLS) with instrumental variables by [5], which we refer to for a detailed data description. The analysis is conducted on a sample of size $n = 5390$, formed by male citizens resident in Austria, Belgium, Czech Republic, Denmark, France, Germany, Italy, Netherland e Sweden. We handle the outcome variable on a logarithmic scale and model both the outcome and the treatment as Gaussian.

Results, displayed in Table 1, show comparable estimates of the linear and two stage methods, yet suggest that each additional year of education respectively lead to an average income increase of 4.2% and 5.2%. The reported mixture model, selected as optimal by the Bayesian Information Criterion, combines two imbalanced components ($\hat{\pi}_1 = 0.075$, $\hat{\pi}_2 = 0.925$). The prevalent class is associated with parameter estimates very close to the OLS solution. Conversely, the smaller group is characterized by a treatment effect remarkably higher ($\hat{\beta}_{11} = 0.157$), thus leading to an ATE eventually comparable with 2SLS estimates in [5], and resulting from the inclusion of instrumental variables.

4 Discussion

When relevant variables are omitted, preventing the interpretation of a regression coefficient in terms of ATE in the presence of selection bias, the proposed model (2) comes to the aid of identifying and accounting for possible sources of heterogeneity, hence reducing distortion and strengthening the CIA.

A notable advantage of the proposed model is its flexibility, both methodologically and interpretatively. Methodologically, the model, detailed with reference to normal linear models and binary treatment, actually covers a much wider range of situations, as it can be readily extended to the use of generalized linear models and treatment belonging to the exponential dispersion family. With

¹ Mostly funded by the European Commission through the 5th, 6th, and 7th framework programmes: projects QLK6-CT-2001-00360, SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5-CT-2005-028857, SHARELIFE, CIT4-CT-2006-028812, SHARE-PREP, 211909, SHARE-LEAP, 227822.

Table 1. Estimated coefficients of models on SHARE data via OLS, 2SLS and mixture with two components

	OLS	2SLS	Mixture		
			Class 1	Class 2	Average
<i>(Intercept)</i>	9.286***	9.183***	5.390***	9.384***	9.137
Years of education	0.042***	0.052.	0.157***	0.045***	0.052
<i>Rural</i>	−0.041*	−0.032	0.291*	−0.049	−0.028
<i>Cohort</i>	0.004***	0.003	−0.000	−0.001	−0.001
SE	0.142***	0.153**	1.239**	0.141**	0.209
DK	0.019	0.018	1.265**	0.019	0.096
DEW	0.094*	0.085.	1.130**	0.092*	0.156
NL	−0.053	−0.045	0.886**	0.001	0.055
FR	−0.165***	−0.159***	1.284**	0.216***	0.282
AT	−0.041	−0.026	0.381	0.103*	0.120
IT	−0.183***	−0.150	1.398**	−0.115	−0.021
CZ	−0.549***	−0.553***	−0.458	−0.571**	−0.564

some primarily computational complications, further flexibility can be achieved through nonlinear modeling and the introduction of non-parametric methods, or the use of recent variable selection techniques in high-dimensional data contexts. Interpretatively, the model allows for the estimation and testing of potential heterogeneity in the effect, as no constraint of equality of effects among the identified groups is specified.

It should be noted that while the natural application of the proposed methodology concerns observational studies, where the presence of heterogeneity resulting in selection bias is expected, another aspect of interest concerns the opportunity, in principle, to use the proposed model even within a randomized experiment or in situations where the CIA is deemed satisfied. Indeed, the proposed model includes as a special case (when $G = 1$) a classic regression model characterized by endogenous treatment and therefore, theoretically, it cannot lead to worsened distortions in the evaluation of the effect compared to what would be obtained with a classic OLS estimation.

References

1. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press (2010)

2. Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E.: Model-Based Clustering and Classification for Data Science with Applications in R. Cambridge University Press (2019)

3. Faria, S., Soromenho, G.: Fitting mixtures of linear regressions. J. Stat. Comp. Sim. **80**(2), 201–225 (2010)

4. Hasan, A., Zhiyu, W., Mahani, A.S.: Fast estimation of multinomial logit models: R package mnlogit. arXiv preprint [arXiv:1404.3177](https://arxiv.org/abs/1404.3177) (2014)

5. Brunello, G., Weber, G., Weiss, C.T.: Books are forever: early life conditions, education and lifetime earnings in Europe. Econ. J. **127**(600), 271–296 (2017)



A Flexible Hybrid Method for Sample Size Calculations in Comparative Poisson Trials

Susanna Gentile^(✉) and Valeria Sambucini

Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
{susanna.gentile, valeria.sambucini}@uniroma1.it

Abstract. Hybrid classical-Bayesian approaches for sample size calculations are used to formally integrate prior beliefs or uncertainty about unknown parameters involved in the traditional power. In this paper, we consider a hybrid procedure that exploits different prior distributions according to the role played by the unknown quantities. The approach is applied to obtain the optimal sample sizes of comparative studies based on Poisson data by using the rate ratio as the measure of interest. A numerical example is provided, allowing the comparison between classical and hybrid methods.

Keywords: gamma prior distributions · hybrid classical-Bayesian approach · Poisson clinical trials · sample size determination

1 Introduction

In clinical trials, the most commonly used approach for sample size determination (SSD) relies on classical hypothesis-based methods of power analysis. The aim is to select the smallest sample size that allows to achieve a desired level for the probability of correctly rejecting the null hypothesis, when the parameter of interest is equal to a fixed *design value* that belongs to the alternative hypothesis. The design value plays a key role in the procedure and is selected as the minimal clinically relevant effect that researchers are interested in detecting.

A hybrid classical-Bayesian approach can be used to formally integrate uncertainty on the design value, maintaining a frequentist framework for the final analysis (see [1]). The SSD criterion is based on the expectation of the classical power function with respect to a *design prior distribution*, that represents an instrumental tool to define a design scenario where the alternative hypothesis is true. This quantity provides a predictive version of the classical power (see [2] and [3]). Moreover, the variance of the test statistic used to construct the rejection region of the null hypothesis typically depends on unknown quantities that are replaced by preliminary estimates. To add flexibility to the SSD method, it is possible to introduce prior distributions, that account for variability around these prior estimates, and to use them to further average the predictive power.

This procedure has been recently proposed (see [4]) with applications to comparative studies based on a binary endpoint.

In this paper, we focus on two-arm clinical trials based on Poisson data and consider the rate ratio as the measure that enables the comparison between treatments. Among several methods proposed to test such parameter (see Gu et al. [5]), we exploit asymptotic tests based on normal approximations. The aim is to illustrate how to implement the hybrid SSD method mentioned above, that takes into account the different role of unknown parameters in the statistical power.

The outline of this paper is as follows. In Sect. 2, we formalize the trial setting and describe the classical SSD procedure for comparative trials based on Poisson data. Section 3 illustrates the hybrid classical-Bayesian method, while in Sect. 4 we present an application based on a numerical example taken from the literature. Finally, Sect. 5 contains some concluding remarks.

2 Classical SSD Method for Comparative Poisson Trials

Let us focus on a randomized two-arm study aimed at comparing an experimental drug (Treatment 1) with a standard one (Treatment 2) by using a count variable as primary endpoint. We consider two independent samples of sizes n_1 and n_2 randomly selected from Poisson populations of rates θ_1 and θ_2 , respectively. In this situation, the natural metric for between treatment comparison is the rate ratio, $RR = \frac{\theta_1}{\theta_2}$, and we assume as parameter of interest its logarithmic transformation, i.e. $\theta = \log(RR)$. Assuming that the counting concerns a “not desirable” event for patients, to establish the superiority of the experimental treatment we consider the following hypotheses $H_0 : \theta \geq \theta_S$ and $H_1 : \theta < \theta_S$, where $\theta_S \leq 0$ denotes the superiority margin.

The MLE of RR is $\widehat{RR} = \frac{\hat{\theta}_1}{\hat{\theta}_2}$, where $\hat{\theta}_j$ is the estimated rate in the j th arm of n_j subjects. We specify in advance the treatment allocation ratio, $r = n_1/n_2$, so that $n_1 = rn_2$. Moreover, by applying the Delta Method, we have that the sampling distribution of the MLE of θ , $Y_{n_2} = \log \widehat{RR}$, is asymptotically approximated by a normal distribution (see Lachin [6])

$$Y_{n_2}|\theta \approx N\left(\theta, \frac{\tau^2(\theta_1, \theta_2)}{n_2}\right), \quad (1)$$

where $\tau(\theta_1, \theta_2) = \sqrt{\frac{1}{r\theta_1} + \frac{1}{\theta_2}}$. The result in (1) is still valid if we replace the asymptotic variance with its plug-in estimate and, thus, standard procedures to make inference on θ exploit the asymptotically normal sampling distribution of Y_{n_2} , denoted by $f(y_{n_2}|\theta)$, whose variance $\frac{\hat{\tau}}{n_2}$ is treated as a known quantity with $\hat{\tau} = \tau(\hat{\theta}_1, \hat{\theta}_2)$.

The typical frequentist SSD procedure takes into account the power function, for a fixed type I error rate, obtained under the assumption that θ is equal to a suitable *design value*, θ^D , which belongs to the alternative hypothesis, i.e.

$$\eta_F^C(\theta^D) = \mathbb{P}_{f(\cdot|\theta^D)}\left(\frac{Y_{n_2} - \theta_S}{\hat{\tau}/\sqrt{n_2}} < z_\alpha\right) = \Phi\left[z_\alpha + \frac{\theta_S - \theta^D}{\hat{\tau}/\sqrt{n_2}}\right], \quad (2)$$

where $\mathbb{P}_{f(\cdot|\theta^D)}$ is the probability measure associated with the asymptotic sampling distribution of Y_{n_2} for $\theta = \theta^D$, while z_α and $\Phi(\cdot)$ denote the α -quantile and the cdf of the standard normal distribution, respectively.

The quantity in (2) is referred to as the *frequentist conditional power* and represents the probability of correctly rejecting H_0 using a frequentist framework to make the decision, when the true value of θ is equal to $\theta^D \in H_1$, which plays the role of the clinical relevant value important to detect. Thus, for a fixed probability threshold $1 - \beta$, the SSD criterion selects the minimum n_2 that satisfies $\eta_F^C(\theta^D) > 1 - \beta$, that is $n_2^{FC} = (z_{1-\beta} - z_\alpha)^2 \hat{\tau}^2 / (\theta_S - \theta^D)^2$.

The corresponding optimal value for n_1 is $n_1^{FC} = rn_2^{FC}$.

3 A Hybrid SSD Method for Comparative Poisson Trials

The traditional approach, based on the frequentist conditional power, critically depends on the choice of the point alternative θ^D and suffers from local optimality. To avoid this drawback, a typical hybrid approach entails the elicitation of a prior distribution, called the *design prior*, that formalizes uncertainty around θ^D . More specifically, we assume $\pi^D(\theta) = N(\theta; \theta^D, \hat{\tau}^2/n^D)$, where the prior sample size n^D can be chosen to regulate the variability around the clinically relevant effect. By averaging the classical power with respect to $\pi^D(\theta)$, we obtain the *frequentist predictive power* (see [2] and [3])

$$\eta_F^P(\pi^D) = \mathbb{P}_{m^D(\cdot)}\left(\frac{Y_{n_2} - \theta_S}{\hat{\tau}/\sqrt{n_2}} < z_\alpha\right) = \Phi\left[\frac{\sqrt{n^D}}{\sqrt{n^D + n_2}}\left(z_\alpha + \frac{\theta_S - \theta^D}{\hat{\tau}/\sqrt{n_2}}\right)\right], \quad (3)$$

where $\mathbb{P}_{m^D(\cdot)}$ denotes the probability measure associated with the marginal distribution of Y_{n_2} , that is $m^D(y_{n_2}) = N\left(y_{n_2}; \theta^D, \hat{\tau}^2\left(\frac{1}{n_2} + \frac{1}{n^D}\right)\right)$.

The power function in (3) provides the probability of making a correct decision by rejecting H_0 , when θ actually belongs to the alternative hypothesis, where we assume that it is distributed according to the design prior. Thus, this latter distribution has to be proper and must support the subspace under H_1 . The optimal value of n_2 , denoted by n_2^{FP} , is selected as the minimum value such that $\eta_F^P(\pi^D) > 1 - \beta$, but a closed-form formula cannot be obtained and a numerical procedure has to be used. Finally, let us notice that the conditional power can be obtained as a special case of the predictive one, when $\pi^D(\theta)$ is chosen as a point-mass distribution centred on θ^D .

3.1 Adding Uncertainty on the Estimate of the Asymptotic Variance

A drawback of the frequentist predictive power consists in the use of a point estimate for the asymptotic variance of Y_{n_2} . This latter quantity, in fact, depends on the unknown Poisson rates, θ_1 and θ_2 , which are replaced by crude estimates. To add flexibility to the procedure, we follow the approach suggested in [4] that accounts for variability around $\hat{\theta}_1$ and $\hat{\theta}_2$. More specifically, we consider two independent gamma prior distributions, $\pi_j(\theta_j) = \text{gamma}(\theta_j; a_j, b_j)$, for $j = 1, 2$. They are used to obtain the *average frequentist predictive power* given by the weighted average of the power in (3),

$$\eta_{AF}^P(\pi^D, \pi_1, \pi_2) = \int_0^\infty \int_0^\infty \eta_F^P(\pi^D, \tau(\theta_1, \theta_2)) \pi_1(\theta_1) \pi_2(\theta_2) d\theta_1 d\theta_2, \quad (4)$$

where

$$\eta_F^P(\pi^D, \tau(\theta_1, \theta_2)) = \Phi \left[\frac{\sqrt{n^D}}{\sqrt{n^D + n_2}} \left(z_\alpha + \frac{\theta_S - \theta^D}{\tau(\theta_1, \theta_2)/\sqrt{n_2}} \right) \right]$$

represents the version of the frequentist predictive power that does not depend on fixed estimates of θ_1 and θ_2 . Note that the power in (4) can be calculated through simulation tools and, for a fixed level $1 - \beta$, the corresponding optimal sample size n_2^{AFP} can be obtained by exploiting numerical procedures. Finally, it is worth pointing out that $\eta_F^P(\pi^D)$ is a special case of $\eta_{AF}^P(\pi^D, \pi_1, \pi_2)$, when degenerate gamma priors at $\hat{\theta}_1$ and $\hat{\theta}_2$ are used.

4 Application to a Multiple Sclerosis Trial

We now apply the hybrid methodology described in the previous section on a numerical example presented by Friede and Schmidli [7] and inspired by confirmatory trials for multiple sclerosis. The primary endpoint for confirmatory studies in this field is usually the annualized relapse rate.

In line with Schmidli and Friedle, we declare the new treatment superior if we cannot reject $H_0 : \theta = \log RR \geq 0$ at level $\alpha = 0.025$. Furthermore, we set the design value equals to $\theta^D = \log(0.75) = -0.288$, while the preliminary estimates used to compute τ^2 are $\hat{\theta}_1 = 0.857$ and $\hat{\theta}_2 = 0.643$, respectively. Note that the preliminary estimates are such that $\log(\hat{\theta}_1/\hat{\theta}_2) = \theta^D$. The corresponding estimate of τ^2 is $\hat{\tau}^2 = 2.722$. Our aim is to size a balanced randomized trial, i.e. $r = 1$, to ensure that a power level of $1 - \beta = 0.9$ is achieved.

To obtain the predictive power, we need to elicit the design prior distribution $\pi^D(\theta) = N(\theta; \theta^D, \hat{\tau}^2/n^D)$ with mean equal to the fixed design value. In particular, we select n^D so that $\pi^D(\theta)$ assigns a 0.999 probability to the symmetric interval around θ^D $(-0.488, 0.088)$. The obtained value is $n^D = 624$. Regarding the two independent prior distributions $\pi_j(\theta_j) = \text{gamma}(\theta_j; a_j, b_j)$,

for $j = 1, 2$, necessary to implement the average predictive power, we specify their expected values equal to the preliminary estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Then, we regulate the dispersion of the distributions through the choice of the prior sample sizes, n_j^0 . The resulting hyperparameters are $a_j = \hat{\theta}_j n_j^0$ and $b_j = n_j^0$. Specifically, we consider two couples of gamma priors by fixing $n_1^0 = n_2^0$ equal to 10 and 20, respectively, and represent their behaviours in the first column of Fig. 1. In order to explain the effect of the prior distributions of θ_1 and θ_2 on the behaviour of $\eta_{AF}^P(\pi^D, \pi_1, \pi_2)$, in the second column of Fig. 1 we represent the probability distributions of $\tau^2(\theta_1, \theta_2)$ induced by the selected gamma densities and obtained by simulation. Finally, in the third column of the same Figure, we compare the behaviour of the frequentist conditional, the predictive and the average predictive powers as a function of n_2 .

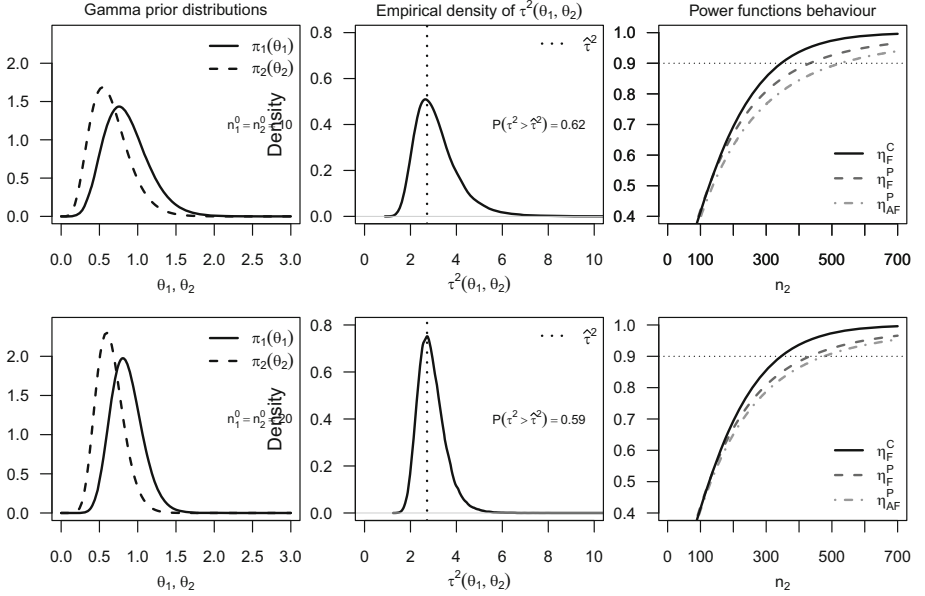


Fig. 1. First column: plot of the prior densities $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$, when $n_1^0 = n_2^0$ are equal to 10 and 20, respectively. Second column: corresponding empirical density of $\tau^2(\theta_1, \theta_2)$. Third column: behaviour of the three power functions $\eta_F^C(\theta^D)$, $\eta_F^P(\pi^D)$ and $\eta_{AF}^P(\pi^D, \tau(\theta_1, \theta_2))$, when $n^D = 624$.

As expected, $\eta_F^P(\pi^D)$ is lower than $\eta_F^C(\theta^D)$ because we are accounting for the uncertainty around the design value θ^D . As for the average predictive power, it rises slower than the standard predictive one in both the cases considered and gets closer to $\eta_F^P(\pi^D)$ as the prior sample sizes $n_1^0 = n_2^0$ increase. This behaviour depends on the distribution of $\tau^2(\theta_1, \theta_2)$ induced by the priors elicited and, consequently, on their impact on the variance of Y_{n_2} . In fact, for these choices of $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$, the estimated probabilities that τ^2 is larger than

$\hat{\tau}^2$ are 0.62 and 0.59, respectively. Consequently, in both cases, the two gamma prior distributions convey jointly the information that, for each value of n_2 , the variance of the test statistic Y_{n_2} is most probably larger than the one estimated using the preliminary estimates. The optimal sample sizes at level $1 - \beta = 0.9$ are consistent with these observations regarding the power functions behaviour. More specifically, the optimal sample size is $n_2^{FC} = 346$ for the conditional frequentist power, $n_2^{FP} = 433$ for the predictive one, $n_2^{AFP} = 526$ for the average predictive power when $n_1^0 = n_2^0 = 10$ and $n_2^{AFP} = 476$ when $n_1^0 = n_2^0 = 20$.

5 Conclusions

Among the possible ways to implement a hybrid method for sample size determination (see [8]), a typical approach is based on the expectation of the traditional power function with respect to a design prior distribution elicited for the clinically relevant effect. This approach takes into account the uncertainty on design expectations, but often exploits preliminary estimates of unknown parameters involved in the final analysis of the trial and fails to account for their variability.

The hybrid SSD method presented in this paper for comparative trials based on Poisson data represents a more flexible procedure in that, instead of using crude prior estimates, it exploits prior distributions for the unknown parameters. Furthermore, it can be easily adapted to studies based on different trial objectives. In fact, without loss in generality, we have considered a superiority trial based on the count of “negative” events for patients. However, the hybrid method we follow can be extended to non-inferiority or equivalence studies and when the hypotheses are reversed because a “positive” event is considered. Finally, here we focus on count data. However, we can adapt this methodology to determine the optimal sample size for randomized controlled trial whenever the sampling distribution of the test statistic is approximately normal (see [4] for examples involving binary data).

References

1. Spiegelhalter, D.J., Abrams, K.R., Myles, J.P.: Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley (2004)
2. Gubbiotti, S., De Santis, F.: Classical and Bayesian power functions: their use in clinical trials. *Biomed. Stat. Clin. Epidemiol.* **2**(3), 201–211 (2008)
3. Sambucini, V.: Bayesian vs frequentist power functions to determine the optimal sample size: testing one sample binomial proportion using exact methods. In: Tejedor, J.P. (ed.) *Bayesian Inference*, pp. 77–97. IntechOpen (2017) <https://doi.org/10.5772/intechopen.70168>
4. Sambucini, V.: Hybrid classical-Bayesian approach to sample size determination for two-arm superiority clinical trials (2024, submitted)
5. Gu, K., Ng, H.K., Tang, M.L., Schucany, W.R.: Testing the ratio of two poisson rates. *Biom. J.* **50**(2), 283–98 (2008). <https://doi.org/10.1002/bimj.200710403>
6. Lachin, J.M.: *Biostatistical Methods: The Assessment of Relative Risks*, 2nd edn. Wiley, Hoboken (2011)

7. Friede, T., Schmidli, H.: Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Stat. Med.* **29**, 1145–1156 (2010). <https://doi.org/10.1002/sim.3861>
8. Kunzmann, K., Grayling, M.J., Lee, K.M., Robertson, D.S., Rufibach, K., Wason, J.M.S.: A review of Bayesian perspectives on sample size derivation for confirmatory trials. *Am. Stat.* **75**(4), 424–432 (2021). <https://doi.org/10.1080/00031305.2021.1901782>



Combining Bootstrap Methods: A Monte Carlo Experiment

Luisa Bisaglia¹ and Margherita Gerolimetto²(✉)

¹ University of Padua, Via Cesare Battisti 241, 35121 Padua, Italy
`luisa.bisaglia@unipd.it`

² Ca' Foscari University Venice, Cannaregio 873, 30121 Venice, Italy
`margherita.gerolimetto@unive.it`

Abstract. The idea of this work is to study whether the combination of different bootstrap methods can lead to an improvement in the performance, as it does in the forecasting framework where is widely used. Given that it represents a rather challenging set-up, we will focus on long memory time series and we will explore the combination of three very well-known bootstrap methods for time series with long range dependence. We present a very preliminary Monte Carlo experiment that provides interesting and promising results.

Keywords: bootstrap · long memory · combination

1 Introduction

Bootstrap methods are still now among the most popular tools to approximate the distribution of certain statistics that would be difficult to be handled otherwise. Bootstrap methods have been proposed in principle for independent observations and afterwards they have been extended to the case when there is serial dependence, such as in time series data. In this context, long memory time series represent quite a complex setting since the strong dependence characterizing those data often requires specific techniques to form a valid resampling strategy.

Traditional bootstrap strategies for time series can be broadly divided into two categories. The first includes methods that resample the residuals of a parametric estimated model (assuming they are *iid*) and the validity of these methods is based on the correct specification of the model and on the consistency of the estimation methods (e.g. [1]). The second category gathers bootstrap techniques that do not strongly rely on parametric restrictions, the most popular in case of long memory time series are the Sieve Bootstrap [2], that is considered semi-parametric, and the Block Bootstrap [3, 4], that is considered nonparametric. The element in common to all those methods is that the final resampling is implemented in approximately independent quantities (either residuals or blocks) that yet is difficult to guarantee in long memory time series whose serial dependence is

often so strong that it persists even when the observations are distant, as emphasized by [5] in a review that presents also an extensive Monte Carlo experiment together with a possible guide for practitioners.

To the best of our knowledge, there is no attempt to consider the possibility of combining bootstrap methods, as it is often done with forecasts. The first contribute to formalize the idea of combining multiple individuals forecasts is by [6]. Still now in the literature it is established that forecast combinations are beneficial. With this in mind, the aim of this work is to investigate whether also for bootstrap methods the idea of combination can lead to a possibly better performance. Given that it represents quite a challenging set-up, we will focus our attention on long memory time series. From the methodological point of view, for now, we will confine ourselves to the Sieve Bootstrap, the Moving Block Bootstrap and the Parametric Bootstrap.

The structure of the paper is as follows. In the second section we present our methodological framework. In the third section we describe the Monte Carlo experiment and some table results.

2 Bootstrap Combination

In this section we will present our methodological framework.

2.1 The Combination of Forecast: State of the Art

The literature on forecast combinations has notably flourished (among others see [7]) for a very recent and rich review). Combining multiple forecasts is now widely adopted to improve the performance. There are several reasons why the forecast combinations perform well, the most intuitive one being that the combination is capable of integrating different information (not completely overlapping) provided by the individual forecasts. The advantage from forecast combination does not depend only on the quality of the individual forecast to be combined but also on how the weights assigned to each forecast are determined. There is a big variety of combination schemes that range from simple combinations that avoid weight estimation (e.g. [8]) to very sophisticated methods that assign specific weights to each individual forecast (e.g. [9]). In this large set of options, equally weighted averages, which do not account for past information regarding the precision of the individual forecasts and correlation between forecast errors, work reasonably well compared to potentially more rigorous combination schemes.

In this paper we will borrow the set-up and the literature advancement in the forecast combination to extend it to bootstrap resampling strategies, as will be presented in the next sections.

2.2 Long Memory Time Series

We will consider a linear process of the form

$$y_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j} \quad (1)$$

where ϵ_t are *iid* variables with zero mean and finite variance and a_j are real-valued constants. When $\sum_{j=0}^{\infty} a_j^2 < \infty$ the process is stationary. Long memory time series are characterized by a significant dependence between distant observation such that the autocovariance function satisfies:

$$\Gamma_y(k) = \text{Cov}(y_t, y_{t-k}) \sim Gk^{2d-1} \quad k \rightarrow \infty \quad (2)$$

where G is a finite constant and d is the memory parameter that governs the strong persistence of y_t . When $|d| < 1/2$ the series is stationary, but it shows long memory (the autocovariances are not summable) and the spectral density (for C positive constant) behaves as $f_y(\lambda) \sim C\lambda^{-2d}$ when $\lambda \rightarrow 0$. One of the best know form of long memory is fractional integration, denoted by $y_t \sim I(d)$ of which ARFIMA(p,d,q) is undoubtedly the most used.

$$\Phi_p(B)(1-B)^d y_t = \Theta_q(B)\epsilon_t \quad (3)$$

where $\Phi_p(B)$ and $\Theta_q(B)$ are polynomials of order p and q respectively, with all roots outside the unit circle and no roots in common.

As for the estimation of the long memory parameter d , the literature is rich of proposals, we will focus here on the most traditional ones, implemented in several packages (for example in R). They are the Log-periodogram estimator (LPE) proposed by [10] and further developed by [11] and the Local Wittle estimator (LWE), proposed by [12] and also further developed by [11]. Due to lack of space, we refer to the original literature for the technical details of the estimators.

2.3 The Combination of Bootstraps

Following the intuition behind the combination of forecasts briefly recalled before, here we present our (very preliminary) proposal of a Combined Bootstrap focussing on the specific case of stationary long memory time series.

It is given a (stationary long memory time series) y_t , $t = 1, \dots, T$ and a statistic S (in our case LPE or LWE) whose distribution should be approximated via bootstrap. The idea is to obtain a combined bootstrap sample y_t^{**} by linearly combining (with a simple combination scheme, for now) the bootstrap samples outcoming from each of three methods suitable for long memory, specifically Parametric Bootstrap, Sieve Bootstrap, Moving Block Bootstrap. The Parametric Bootstrap [1] moves from the assumption that the time series y_t , $t = 1, \dots, T$ is the realization of an *ARFIMA*(p, d, q) process (3). In very simple terms, once the model is estimated, y_t^* will be computed from the estimated model using simulated normal random variables (with mean and variance obtained from the parametric model) in place of the residuals and the unknown parameters replaced by the estimated values. The Sieve Bootstrap [2] is a semiparametric technique and it is based on the approximation of the data generating process with an *AR*(h) model. The bootstrap samples are drawn from the studentized residuals obtained from the autoregressive approximation. The Block Bootstrap is an

entirely nonparametric strategy that is based on resampling block of observations that can be overlapping [4] or not [3]. The approximate independence of the block is necessary for this bootstrap being successful and in case of long memory this is very difficult to find [13]. [14] recommend in this case the use of overlapping blocks (so called Moving Block Bootstrap) and provide some rules of thumb for the blocks length.

The statistic S is evaluated on the Combined Bootstrap sample. The algorithm for $b = 1, \dots, B$ bootstrap samples is the following:

Step 1: Resample y_t with the Parametric Bootstrap and obtain $y_{b,t}^{*PB}$

Step 2: Resample y_t with the Sieve Bootstrap and obtain $y_{b,t}^{*SB}$

Step 3: Resample y_t with Moving Block Bootstrap and $y_{b,t}^{*MBB}$

Step 4: Combine (with a simple linear scheme) the three bootstrap samples and obtain $y_{b,t}^{**}$:

$$y_{b,t}^{**} = \frac{y_{b,t}^{*PB} + y_{b,t}^{*SB} + y_{b,t}^{*MBB}}{3} \quad (4)$$

Step 5: Repeat steps 1–4 to obtain B bootstrap samples and calculate B values of the desired statistic S_b^{**}

Step 6: Approximate the distribution of the statistic S with the bootstrap distribution (where I is the indicator function):

$$FS^{**}(s) = \frac{1}{B} \sum_{b=1}^B I(S_b^{**} \leq s)$$

As mentioned, this bootstrap algorithm is at the moment very preliminary. Following the hints coming from the literature on combination of forecasts, we begin here with a simple linear combination scheme that combines the most suitable time domain tools for bootstrapping long memory time series, with the intention to put forward this idea via possibly more complex and sophisticated weighting scheme and through the consideration of other bootstrapping methods.

3 Monte Carlo Experiment

In order to investigate the performance of the Combined Bootstrap with respect to the other bootstrap methods we intend to carry out a Monte Carlo experiment. Here we will present some preliminary results that will be the guidance for a future more extensive simulation design.

Along the lines of the Monte Carlo experiment in [5], for now, we consider $S = 1000$ Monte Carlo simulations from an $ARFIMA(0, d, 0)$ model, for $T = 250, 500$ sample sizes and for $d = 0.2, 0.4$ values of the long memory parameter. The statistics under examinations are the estimators LWE and LPE of the long memory parameter where $m_{LWE} = 1 + T^{0.65}$ and $m_{LPE} = 1 + T^{0.8}$ (following the related literature and [5]). For the implementation of the Parametric Bootstrap, an $ARFIMA(0, d, 0)$ model is estimated via maximum likelihood. As for the Sieve Bootstrap, the autoregression is estimated via maximum likelihood and

the order of the model is selected via AIC minimization. Finally, for the Moving Block Bootstrap the block length is $k = \lceil T^{1/2} \rceil$, as in [14]. The simulations are run with the software R [15].

The performance is expressed with two measures [5]. The first is the Root Mean Square Deviance (RMSD) of the bootstrap probability density function to the true probability function (obtained via Monte Carlo):

$$RMSD(boot) = \sqrt{\frac{1}{S} \sum_{j=1}^S \left(p_{boot}(\hat{d}_j - d_0) - p_{mc}(\hat{d}_j - d_0) \right)^2} \quad (5)$$

where *boot* is one of the bootstrap methods, d_0 is the true value, $p_{mc}(\hat{d}_j - d_0)$ is the ordinate of a kernel density based on the S estimates $\hat{d}_j - d_0$, $j = 1, \dots, S$ (evaluated on $\hat{d}_j - d_0$), $p_{boot}(\hat{d}_j - d_0)$ is average over the S Monte Carlo simulations of the kernel density based on $\hat{d}_j^b - \hat{d}_j$ the estimates from the $b = 1, \dots, B$ bootstrap samples (evaluated on $\hat{d}_j - d_0$). The closer is the RMSD to zero, the better the performance of the bootstrap strategy. The second measure of performance is the coverage frequency of confidence intervals. The percentiles coming from any of the bootstrap strategy are constructed in general as follows: $CI_{1-\alpha}^*(\hat{d}) = (\hat{d}_{\alpha/2}; \hat{d}_{1-\alpha/2})$ where $\hat{d}_\alpha$ is the $B\alpha$ ordered value of the bootstrap estimate of d with, in turn LWE and LPE. The closer the coverage frequency to $1 - \alpha$ (for us 95%), the better the performance of the bootstrap.

In Table 1 are provided some of the results of the preliminary experiment. We propose the results for $d = 0.4$ and $T = 500$, the others are available upon request, but their meaning is substantially analogous.

Table 1. RMDS and Coverage Frequency of bootstrap distributions for LWE and LPE

LWE			LPE		
Bootstrap	RMSD	Cov Freq	Bootstrap	RMSD	Cov Freq
Comb Boot	0.825	0.901	Comb Boot	1.802	0.989
SB Boot	1.058	0.831	SB Boot	1.759	0.972
MBB Boot	1.317	0.740	MBB Boot	2.311	0.944
PB Boot	1.468	0.991	PB Boot	2.150	0.994

To describe these preliminary results we need first to clarify that the data generating process considered in this experiment is the same as the model estimated in the Parametric Bootstrap algorithm, meaning that the Parametric Bootstrap is implemented in the best scenario and is expected to perform at best. We did this on purpose, so to have some kind of upper bound of the ideal performance with which the other bootstraps can be compared. As we can see, the Combined Bootstrap has a performance comparable with the parametric bootstrap in terms

of coverage frequency and even better than the Parametric Bootstrap in terms of RMSD. The intuition we get from these set of preliminary results is that the Combined Bootstrap, if developed with possibly more sophisticated tools (such as a different weighting scheme for the linear combination or a more wide and informed selection of the component bootstraps) could significantly represent a valid alternative to existing bootstrap strategy.

References

1. Andrews, D.W.K., Lieberman, O., Marmer, V.: Higher-order improvements of the parametric bootstrap for long-memory Gaussian processes. *J. Econometrics* **133**, 673–702 (2024)
2. Buhlmann, P.: Sieve bootstrap for time series. *Bernoulli* **3**, 123–148 (1997)
3. Carlstein, E.: The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Stat.* **14**, 1171–1179 (1986)
4. Kunsch, H.R.: The Jackknife and bootstrap for general stationary observations. *Ann. Stat.* **17**, 1217–1261 (1989)
5. Artech, J.: Bootstrapping long memory time series: application in low frequency estimators. *Econometrics Stat.* **29**, 1–15 (2024)
6. Bates, J.M., Granger, C.W.J.: The combination of forecasts. *OR* **2**, 451–68 (1969)
7. Wang, X., Hyndman, R., Li, F., Kang, Y.: Forecast combinations: an over 50-year review. *Int. J. Forecast.* **39**, 1518–1547 (2023)
8. Petropoulos, F., Svetunkov, I.: A simple combination of univariate models. *Int. J. Forecast.* **36**, 110–115 (2020)
9. Wang, X., Kang, Y., Petropoulos, F., Li, F.: The uncertainty estimation of feature-based forecast combinations. *J. Oper. Res. Soc.* **73**, 979–993 (2022)
10. Geweke, J., Porter-Hudack, S.: The estimation and application of long-memory time series models. *J. Time Ser. Anal.* **4**, 221–237 (1983)
11. Robinson, P.M.: Log-periodogram regression of time series with long-range dependence. *Ann. Stat.* **23**, 1048–1072 (1995)
12. Fox, R., Taqqu, M.S.: Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Stat.* **14**, 517–532 (1986)
13. Lahiri, S.N.: On the moving block bootstrap under long range dependence. *Statist. Probab. Lett.* **18**, 405–413 (1993)
14. Kim, Y.M., Nordman, D.J.: Properties of a block bootstrap under long-range dependence. *Sankhya Indian J. Stat.* **73-A**, 79–109 (2011)
15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023)



Microbiome Enterotype Detection via a Latent Variable Allocation Model

Alice Giampino^(✉), Roberto Ascari, and Sonia Migliorati

Department of Economics Management and Statistics, University of Milano-Bicocca,
Milan, Italy

alice.giampino@unimib.it

Abstract. Human gut microbiome studies are of critical importance to evaluate the health of an individual. Understanding the interactions within microbial communities has a crucial role in comprehending human biological systems. Topic modelling techniques are usually employed to analyze textual documents and exploit a latent variable framework to find hidden topics in the data. Quite interestingly, a clear semantic correspondence between text and microbiome analyses allows applying these techniques to detect enterotypes. One of the most commonly used models in the field is the Latent Dirichlet Allocation which, however, suffers from some limitations due to the stiffness of its standard topic prior distribution, the Dirichlet. This study proposes a flexible Dirichlet distribution as topic prior distribution in the context of microbial systems. The goal is to detect the important taxa characterizing different enterotypes.

Keywords: Dirichlet distribution · topic modeling · microbiome · Gibbs sampling

1 Introduction

The human gut microbiome is the collection of genes associated with the community of microorganisms, i.e. the microbiota, living in our digestive tract. It exhibits significant variability among individuals, and it is influenced by several factors, such as genetic differences among human populations, dietary habits, etc. [1]. Kurokawa et al. [6] underscore this variability, highlighting how factors such as genetics contribute to the distinct microbial composition observed across various human groups. Additionally, studies such as the one conducted by Mai and McCrary [7] emphasize the role of dietary choices in shaping the gut microbiome, with different dietary habits leading to unique microbial profiles. Topic modeling techniques leverage a latent variable framework to uncover hidden thematic patterns within data, often applied to find topics within textual documents. In the context of microbiology, particularly in studying the human microbiome, these approaches find relevance in detecting the complex interactions among microbial taxa, which collectively form the microbiota, as they allow the identification of enterotypes, i.e. representative profiles of bacterial compositions showing particular functional associations [4]. These (latent) enterotypes

are linked to individuals' health status and are believed to respond differently to diet and drug intake [12]. Thus, their detection is of particular relevance.

2 Latent Topic Models

Topic modeling techniques work by assuming that each word in a textual document is generated according to one among T topics, where a topic is thought of as a discrete probability distribution over the set of unique words (i.e., a vocabulary $\mathcal{V}$). Thus, a specific document contains words generated by different (latent) topics affecting all the documents in a given corpus. As a consequence, we can represent the d -th document using a vector $\boldsymbol{\theta}_d = (\theta_{d,1}, \dots, \theta_{d,T})^\top$, where $\theta_{d,t}$ indicates the proportion of words in document d generated from topic t , with $d = 1, \dots, D$ and $t = 1, \dots, T$. It is evident that $\boldsymbol{\theta}_d$ lies within the T -part simplex $\mathcal{S}^T = \{\boldsymbol{\theta} : \theta_t > 0, \sum_{t=1}^T \theta_t = 1\}$. Similarly, each topic may be expressed as a vector $\boldsymbol{\phi}_t \in \mathcal{S}^V$ whose generic element represents the probability that topic t assigns to a given word in the vocabulary.

The connection between microbiome data and text documents can be established by using the analogy outlined, among others, by Sankaran et al. [10]. The collection of biological samples may be considered as a corpus, with each patient/biological sample representing one textual document. Furthermore, the biological taxa may be considered as the unique words forming the vocabulary. In this framework, the latent topics may be considered as latent enterotypes.

Among the different latent topics models employed in this field, the latent Dirichlet allocation (LDA, [3]) stands out as one of the most widely applied. The LDA assumes that both the parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ follow a Dirichlet distribution defined on the proper simplex space, namely

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \in \mathbb{R}_+^T \quad \text{and} \quad \boldsymbol{\phi}_t \sim \text{Dir}(\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \mathbb{R}_+^V.$$

Although its popularity, the LDA suffers from some limitations due to Dirichlet's restrictive parameterization of the covariance matrix. Indeed, once its barycenter has been fixed, the Dirichlet deserves just one parameter (i.e., $\boldsymbol{\alpha}^+ = \sum_{t=1}^T \boldsymbol{\alpha}_t$) to model the entire covariance matrix.

In the LDA model, Bayesian inference is typically carried out relying on two different methods: the variational Bayes (VB) and the collapsed Gibbs sampling (CGS) scheme. The VB is known for its fast convergence and scalability to large datasets. However, it often relies on simplifications and approximations, which can result in biased estimates or inaccurate posterior distributions. Standard Gibbs sampling (GS) techniques are largely used for their convergence to the real posterior distribution but may suffer from being computationally intensive and slow, particularly when dealing with large datasets. The CGS differs from standard GS because it marginalizes some parameters (i.e., the $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$'s) out from the joint posterior distribution; these parameters are later estimated relying on some conjugacy property.

The introduction of a more flexible latent topic model may help in analyzing real microbiome data and detecting latent microbial enterotypes having a biological interpretation.

2.1 Flexible Latent Dirichlet Allocation

A generalization of the LDA, namely the flexible LDA (FLDA) is obtained by assuming a flexible Dirichlet distribution (FD, [8, 9]) for each θ_d . The FD is an identifiable finite mixture model with Dirichlet components defined as

$$\text{FD}(\theta; \alpha, \tau, \mathbf{p}) = \sum_{t=1}^T p_t \text{Dir}(\theta; \alpha + \tau \cdot \mathbf{e}_t),$$

where $\mathbf{p} \in \mathcal{S}^T$ is a vector of mixing weights, $\tau > 0$, and $\mathbf{e}_t$ represents the null vector with the t -th element equal to 1. On the one hand, this mixture structure introduces additional parameters, relaxing the constraints imposed on the covariance matrix. On the other hand, it maintains some statistical properties of the Dirichlet, such as the conjugacy with respect to the multinomial sampling scheme. It is worth noting that the FD incorporates the Dirichlet distribution as a special case when $\tau = 1$ and $p_t = \alpha_t / (\sum_{l=1}^T \alpha_l)$ for $t = 1, \dots, T$. Hence, the FLDA model includes the LDA as a special case.

We generalize the approach of Griffith and Steyvers [5] and sketch a CGS algorithm for the FLDA model. In particular, we aim at obtaining estimates for the FLDA parameters, namely $\theta_1, \dots, \theta_D$ and $\phi_1, \dots, \phi_T$. The enterotype label of the n -th bacterial read ($n = 1, \dots, N_d$) in the d -th biological sample, is introduced as a latent discrete random variable $Z_{d,n}$. The probability that a bacterial read is generated from enterotype t given the other enterotypes assignments $\mathbf{z}_{-(d,n)}$ represents the full conditional distribution and is specified as

$$p_Z(t) \equiv p(Z_{d,n} = t | \mathbf{z}_{-(d,n)}, \mathcal{C}, \alpha, \tau, \mathbf{p}, \beta) \propto \frac{(\alpha_t + c_{t,d,\cdot}^-) (\beta_{v_{d,n}} + c_{t,\cdot,w_{d,n}}^-)}{(\beta^+ + c_{t,\cdot,\cdot}^-)} \cdot \left\{ \sum_{h=1}^T p_{d,h}^* + p_{d,t}^* \left(\frac{\tau_t}{\alpha_t + c_{t,d,\cdot}^-} \right) \right\},$$

$t = 1, \dots, T$, where $p_{d,t}^* = p_t \frac{(\alpha_t + \tau)^{[c_{t,d,\cdot}^-]}}{(\alpha_t)^{[c_{t,d,\cdot}^-]}}$, $w_{d,n} \in \mathcal{V}$ indicates which taxon is associated with the n -th bacterial read in the d -th biological sample, $x^{[n]} = x(x+1) \cdot \dots \cdot (x+n-1)$ denotes the rising factorial function, and $c_{t,d,\cdot}$, $c_{t,\cdot,w}$, and $c_{t,d,\cdot}$ are defined as the summation over the proper index of the counts $c_{t,d,v} = \sum_{n=1}^{N_d} \mathbb{I}(z_{d,n} = t, w_{d,n} = v)$, the latter representing the number of times that taxon v is assigned to enterotype t in the biological sample d .

The CGS algorithm exploits the full conditional distributions, and it consists of the following steps:

1. Initialize the vector $\mathbf{z}$ (e.g., randomly) and compute the counts $c_{t,d,v}^{(0)}$;
2. For $b = 1, \dots, B$:
 - For each bacterial read $w_{d,n}$, $d = 1, \dots, D$ and $n = 1, \dots, N_d$:
 - sample a new enterotype label $z_{d,n}^{(b)}$ for $w_{d,n}$ from $p_Z(\cdot)$;
 - update the counts $c_{t,d,v}^{(b)}$.

- Use $\mathbf{z}^{(b)}$ to compute the estimates $\hat{\boldsymbol{\theta}}_d^{(b)}$ and $\hat{\boldsymbol{\phi}}_t^{(b)}$.

Estimates of $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ are obtained by taking advantage of the conjugacy property of the FD and Dirichlet distributions with respect to the multinomial sampling scheme[2]. In particular, by having the enterotype labels $\mathbf{z}^{(b)}$, $b = 1, \dots, B$, we can estimate the parameters through their posterior mean [2]:

$$\hat{\boldsymbol{\theta}}_d^{(b)} = \frac{\boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau \mathbf{P}_d^{*(b)} / p_{d,+}^{(b)}}{\boldsymbol{\alpha}^+ + \tau + N_d} \quad \text{and} \quad \hat{\boldsymbol{\phi}}_t^{(b)} = \frac{\boldsymbol{\beta} + \mathbf{c}_t^{(b)}}{\boldsymbol{\beta}^+ + c_{t,.,}^{(b)}},$$

where $p_{d,+}^{(b)} = \sum_{h=1}^T p_{d,h}^{*(b)}$, $\mathbf{c}_d^{(b)} = (c_{1,d,.,}^{(b)}, \dots, c_{T,d,.,}^{(b)})^\top$, and $\mathbf{c}_t^{(b)} = (c_{t,.,1}^{(b)}, \dots, c_{t,.,V}^{(b)})^\top$.

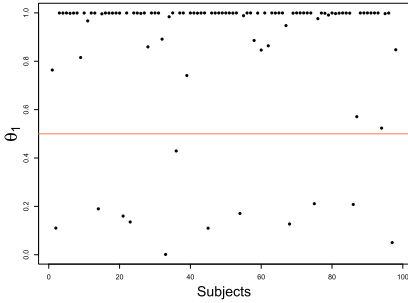
3 Human Gut Microbiome Data

The COMBO dataset contains information on the human gut microbiome, and it was initially introduced by Wu et al. [12]. Data refer to a cross-sectional examination that involved 98 healthy volunteers, consisting of the extraction of the DNA from stool samples underwent 454/Roche pyrosequencing of 16S rRNA gene segments. In this analysis, we consider the counts from three different taxa, namely Bacteroides, Prevotella, and Ruminococcus. This choice is quite widespread (e.g., see [1, 11], and [12]) since these genera are crucial for the detection of enterotypes which, in turn, are strictly linked to health status and vitamin production [12]. In particular, Arumugam et al. [1] define three enterotypes: the first enterotype being affected by the Bacteroides taxon, the second by Prevotella, and the third by Ruminococcus. Wu et al. [12] found out that the COMBO dataset is affected by just the first and the second enterotypes. Thus, we consider a vocabulary $\mathcal{V}$ with $V = 3$ taxa and a corpus composed of $D = 98$ volunteers/biological samples, with a total of 433643 bacterial reads. We then look for $T = 2$ and $T = 3$ latent microbial enterotypes and try to identify which taxa characterize each of them. To this end, we run the FLDA model for $B = 7000$ iterations. Results are reported in Table 1 and Fig. 1.

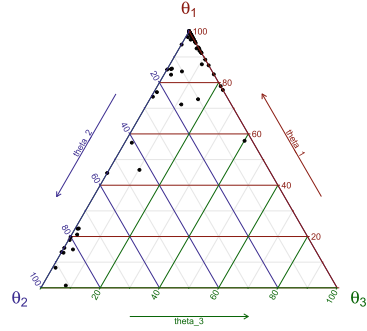
Table 1. Estimated $\boldsymbol{\phi}_t$ parameters for $T = 2$ and $T = 3$ microbial enterotypes.

	$T = 2$		$T = 3$		
Taxa	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Bacteroides	0.9802	0.0370	0.99998	0.0046	0.2633
Prevotella	0.0001	0.9516	0.00001	0.9952	0.0001
Ruminococcus	0.0197	0.0114	0.00001	0.0002	0.7366

Inspecting Table 1, we can see that, with $T = 2$, the FLDA model detects two latent enterotypes, each being influenced by just one taxon. In particular, the first enterotype is mainly composed of the Bacteroides taxon ($\hat{\phi}_{1,1} = 0.9802$),



(a) Two enterotypes.



(b) Three enterotypes.

Fig. 1. Plots of the vectors of estimated microbial enterotypes proportions on the 98 biological samples with different numbers of enterotypes.

whereas the second is represented by a large amount of the Prevotella taxon ($\hat{\phi}_{2,2} = 0.9516$). The same conclusion holds when the model looks for $T = 3$ enterotypes: the first and the second enterotypes are influenced by Bacteroides and Prevotella taxa, respectively, whereas the third assigns a high probability to Ruminococcus taxon.

Figures 1a and 1b show the representation of the 98 volunteers as enterotypes proportion, by plotting the corresponding estimates of $\theta_1, \dots, \theta_D$ for $T = 2$ and $T = 3$, respectively. In particular, with $T = 2$, we can observe that 12 subjects (i.e., those below the horizontal line at 0.5 in Fig. 1a) have a higher predominance of bacteria arising from the second enterotype, whereas the remaining subjects have a microbiome composition mainly affected by the first one. Similar considerations hold also when $T = 3$ (Fig. 1b), where the points in the ternary diagram are shrunk towards two vertices, meaning that the third enterotype (i.e., the only one influenced by the Ruminococcus taxon) has generally little impact on the observed sample.

Interestingly, these results are completely coherent with the biological interpretation described by Arumugam et al. [1] and by Wu et al. [12], thus suggesting that the FLDA is a promising methodological enhancement for microbiome analysis.

References

1. Arumugam, M., et al.: Enterotypes of the human gut microbiome. *Nature* **473**(7346), 174–180 (2011)
2. Ascari, R., Migliorati, S., Ongaro, A.: A new prior distribution on the simplex: the extended flexible Dirichlet. *Book of Short Paper SIS* (2020)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**(4–5), 993–1022 (2003). <https://doi.org/10.1016/b978-0-12-411519-4.00006-9>

4. Deek, R.A., Li, H.: A zero-inflated latent Dirichlet allocation model for microbiome studies. *Front. Genet.* **11**, 602594 (2021)
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proc. Natl. Acad. Sci.* **101**(suppl.1), 5228–5235 (2004)
6. Kurokawa, K., et al.: Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14**(4), 169–181 (2007)
7. Mai, V., McCrary, Q.M., Sinha, R., Gleib, M.: Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: an observational study in African American and Caucasian American volunteers. *Nutr. J.* **8**, 1–10 (2009)
8. Migliorati, S., Ongaro, A., Monti, G.S.: A structured Dirichlet mixture model for compositional data: inferential and applicative issues. *Stat. Comput.* **27**, 963–983 (2017)
9. Ongaro, A., Migliorati, S.: A generalization of the Dirichlet distribution. *J. Multivar. Anal.* **114**, 412–426 (2013)
10. Sankaran, K., Holmes, S.P.: Latent variable modeling for the microbiome. *Biostatistics* **20**(4), 599–614 (2019)
11. Xia, F., Chen, J., Fung, W.K., Li, H.: A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69**(4), 1053–1063 (2013)
12. Wu, G.D., et al.: Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**(6052), 105–108 (2011)



Model-Based Clustering of Spatial Time Series Through the BayesMix library

Matteo Gianella^(✉) and Alessandra Guglielmi

Department of Mathematics, Politecnico di Milano, Milan, Italy
{matteo.gianella,alessandra.guglielmi}@polimi.it

Abstract. In this work, we consider time series of daily concentrations of PM₁₀ monitored in Lombardia and Emilia-Romagna during 2018. With the aim of clustering those spatial time series, we propose a Bayesian nonparametric mixture of autoregressive processes and assume as mixing measure a spatial product partition model. We focus on the implementation of this model into **BayesMix**, a new C++ library for Bayesian inference on nonparametric mixture models via Markov Chain Monte Carlo. The main feature of this library is its extensibility, which guarantees a seamless integration of new classes of mixture models, like the one we introduce in this paper, without compromising efficiency.

Keywords: spatial time series · mixture models · C++ programming

1 Introduction

Air pollution is defined as the emission of gases and particulates into the atmosphere due to human activities and resulting in adverse changes. One of the most harmful gaseous pollutants to human health is particulate matter, or fine dust, which is named according to its size in micrometers, PM₁₀ and PM_{2.5}. These pollutants are monitored on a daily or a weekly basis in georeferenced stations scattered throughout the territory. The data that we are going to consider in this work are weekly concentrations of PM₁₀ in 112 stations in Lombardia and Emilia-Romagna during 2018. We call this type of data *spatial time series*.

In this work, our goal is clustering spatial time series, i.e., the station labels, according to their trends in terms of pollutant concentrations. The resulting partition, in fact, might offer an important insight for a better understanding of the factors that contribute to air pollution. Since data are spatial, we need to consider the underlying spatial structure in the prior for the random partition of stations. In the Bayesian paradigm, clustering can be easily obtained through mixture models, since they allow to decompose complex data-generating processes into simpler components, which are easier to handle. To do so, we model our time series as a mixture of autoregressive processes of first order with cluster-specific innovation parameters. The spatial locations in which the time series have been recorded drives the partitioning mechanism via a spatial product partition model [1].

The model we propose is not completely new in the literature, as it consists of well known components from Bayesian nonparametric literature. However, the main objective of this study is to demonstrate the easy integration of such model into the **BayesMix** library [2]. **BayesMix** is a C++ library developed to provide a self-contained ecosystem to perform inference on nonparametric mixture models in the Bayesian framework via Markov Chain Monte Carlo (MCMC). Even if the core part of the library is written in C++, **BayesMix** provides both a Python and a R interface that allow users to not interact directly with the terminal and use their preferred programming language. The driving idea behind this library is *extensibility*: new classes of mixture models can be seamlessly integrated in this library and used in combination with already available prior distributions and algorithms. Moreover, extensibility does not come with a compromise on the efficiency: in the following we will see in practice how to extend this library and its performance in terms of computational time and quality of inference.

The remainder of this work is organised as follows: In Sect. 2 we provide a full description of the Bayesian model for spatial time series clustering and the key building blocks of **BayesMix**. Then, in Sect. 3 we detail the required steps to implement the specified model in the **BayesMix** ecosystem. Finally, in Sect. 4 we analyse our data and briefly comment on the findings we obtain.

2 Preliminaries

In this section we provide a detailed description of the model and key concepts of the **BayesMix** library.

2.1 A Model for Spatial Time Series Clustering

Consider observations $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$, where $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,T})$ is the time series that collects weekly averages of log-concentrations of PM₁₀ in station i . Each station i is associated to a location $\mathbf{s}_i$, for $i = 1, \dots, n$, where $n = 112$. In time series analysis, data are often modelled via autoregressive processes [3]. After exploratory data analysis, we have considered an autoregressive process of first order. We recall that a vector $\mathbf{y}$ follows an autoregressive process of first order with innovation parameter ρ and variance σ^2 if:

$$\begin{cases} y_t \mid y_{t-1}, \rho, \sigma^2 \sim \mathcal{N}(\rho y_{t-1}, \sigma^2) & t = 2, \dots, T \\ y_1 \sim \mathcal{N}(y_0, \sigma^2) \end{cases},$$

and we write $\mathbf{y} \sim AR_1(\rho, \sigma^2)$.

Since our main goal is clustering stations according to the trend of the PM₁₀ pollutant, we assume each time series to be modelled as a mixture of autoregressive processes of first order. If we introduce the latent cluster allocation variables c_i , one for each observation, the likelihood for our data can be written as:

$$\mathbf{y}_i \mid c_i = h, \boldsymbol{\rho}, \boldsymbol{\sigma}^2 \stackrel{\text{ind}}{\sim} AR_1(\rho_h, \sigma_h^2), \quad i = 1, \dots, n. \quad (1)$$

where $\boldsymbol{\rho} = (\rho_1, \dots, \rho_H)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_H^2)$ are H dimensional vectors collecting the cluster-specific innovation parameters and variances, respectively. Hence, the clustering mechanism is driven by the parameters characterising the autoregressive process, which will be cluster-specific. As a prior for $(\boldsymbol{\rho}, \boldsymbol{\sigma}^2)$, we assume independent Normal-Inverse-Gamma distributions for each component:

$$\rho_h \mid \sigma_h^2 \stackrel{\text{iid}}{\sim} \mathcal{N}(\rho_0, \sigma_h^2/\lambda_0), \quad (2)$$

$$\sigma_h^2 \stackrel{\text{iid}}{\sim} \text{InvGamma}(a_0, b_0). \quad (3)$$

This choice of the prior makes the model inside each cluster conjugate.

Finally, the clustering process is driven by a spatial Product Partition Model (sPPM) [1]. In this way, we directly model the partition of spatial locations into clusters taking into account the spatial location where each time series observation has been recorded. The partition probability function we assume in this model is the following:

$$\Pi_H^n[(n_1, \mathbf{s}_1^*), \dots, (n_H, \mathbf{s}_H^*)] \propto \alpha^H \prod_{h=1}^H \Gamma(n_h) \prod_{i,j \in C_h} \mathbb{I}(\|\mathbf{s}_i - \mathbf{s}_j\| \leq a), \quad (4)$$

where $C_h = \{i : c_i = h\}$ and $\mathbf{s}_h^* = \{\mathbf{s}_i : i \in C_h\}$, for $h = 1, \dots, H$. Notice that (4) is the hard cohesion function introduced in [1]. The spatial prior distribution for the random partition completes our model for spatial clustering.

Since the model within each cluster is conjugate, we rely on Neal's Algorithm 2 [4] to carry out posterior inference. Nevertheless, implementing a marginal sampler for this model from scratch may require a long time. We choose to rely on the **BayesMix** library, which provides a self-contained ecosystem for inference on Bayesian nonparametric mixture models via MCMC. In Sect. 3 we show how to implement an MCMC sampler for model (1)–(4) through **BayesMix**.

2.2 The BayesMix Library

Here we give a general overview of the main building blocks in **BayesMix**. In order to understand the ratio between the different modules that compose **BayesMix**, we start with the definition of a generic mixture model under the Bayesian approach. In general, any mixture model can be written as follows:

$$y_i \mid c_i = h, \boldsymbol{\tau} \stackrel{\text{ind}}{\sim} f(\cdot \mid \tau_h), \quad i = 1, \dots, n \quad (5)$$

$$\tau_h \mid G_0 \stackrel{\text{iid}}{\sim} G_0, \quad h = 1, \dots, H \quad (6)$$

$$c_i \mid \mathbf{w} \stackrel{\text{iid}}{\sim} \text{Categorical}(\{1, \dots, H\}, \mathbf{w}), \quad i = 1, \dots, n \quad (7)$$

$$\mathbf{w} \sim \pi(\mathbf{w}). \quad (8)$$

Note that in model (5)–(8) we have introduced, for inferential purposes, the latent cluster allocation variables c_i , for $i = 1, \dots, n$. The main building modules of **BayesMix** are essentially three:

1. **The Hierarchy module:** those modules encode equations (5)–(6), that is the Bayesian model for the data inside a specific cluster. All likelihoods $f(\cdot | \tau_h)$ are implemented in a **Likelihood** class, while each choice of G_0 is implemented in a **PriorModel** class. The **Hierarchy** class, then, is the composition between a **Likelihood** and a **PriorModel**. The sampling from the full conditional of τ_h is performed in an **Updater** class.
2. **The Mixing module:** those modules encode equations (7)–(8), that is the prior on the mixture weights $\mathbf{w}$ and/or the induced exchangeable partition probability function (EPPF).
3. **The Algorithm module:** the **Algorithm** classes are in charge of running the MCMC simulations. An **Algorithm** operates on a **Mixing** and several **Hierarchies** (or clusters), calling their appropriate update methods (and passing the appropriate data as input).

These are the essential information required to understand how the **BayesMix** library works: in this way, one can understand how such library can be extended to implement an MCMC algorithm to sample from the posterior distribution of the model we have presented here. For those interested in a detailed explanation of the software, including the class hierarchy and the application programming interfaces (APIs) for each class, see [2].

3 Extending the Library

In this section, we show in practice the required step to extend **BayesMix** in order to code the MCMC for approximating the posterior of our spatial time series clustering model.

A new Hierarchy class

In order to implement the cluster-specific model, we need to specify (i) a new **Likelihood** class that implements the autoregressive process of first order and (ii) how to update the associated cluster-specific parameters, through an appropriate **Updater** class. The likelihood for an autoregressive process of first order can be implemented in **BayesMix** with few lines of code. In fact, it is sufficient to specify if it requires covariates and if it is suited for univariate or multivariate data. Lastly, we must provide the computation of the log probability density function (lpdf) specified in (1) with the `compute_lpdf()` method. The prior model in (2)–(3) is already implemented in the library and makes the cluster-specific model conjugate. Then, the implementation of the **Updater** boils down to the computation of the posterior parameters $(\rho_p, \lambda_p, a_p, b_p)$, and this is the only method the user needs to implement in the **BayesMix** framework. Then, the **Hierarchy** class can be implemented by combining the AR_1 likelihood with the Normal-Inverse-Gamma prior model and specifying the conjugate updater. Finally, since the model is conjugate, we must provide the computation of the marginal lpdf.

*A new **Mixing** class*

In order to implement the prior for the random partition in (4), we need to define a new **Mixing** class. In this case the mixing measure is given in terms of the partition probability function; then, it is sufficient to provide the code that computes, for any observation i (i.e., the i -th time series), the prior probability w_h to be assigned to an already existing cluster and the probability w_{H+1} to be assigned to a new cluster $H + 1$. In case we consider the partition probability function in (4), those probabilities are proportional to:

$$w_h \propto n_h \prod_{j \in C_h} \mathbb{I} \{ \|s_i - s_j\| \leq a \} \quad h = 1, \dots, H; \quad w_{H+1} \propto \alpha.$$

These simple class definitions are enough to implement the MCMC algorithm to sample from our model for spatial time series clustering. In fact, since the hierarchy is conjugate, we can rely on Neal’s Algorithm 2, already implemented in **BayesMix**.

4 Performance and Posterior Inference

Coming back to our application, we run our MCMC sampler for a total of 2000 iterations, discarding the first half as burn-in. Prior hyperparameters for the autoregressive parameters are set following an empirical Bayes approach, so that the prior mean and variance are close to the corresponding empirical values. As far as the spatial Product Partition Model is concerned, we fix α in (4) so that, a priori, the expected number of clusters is equal to 3 and a in (4) by looking at all pairwise distances between the locations s_i . In particular, according to these assumptions, we have that $\alpha = 0.5$ and $a = 250$ km. The sampler implemented via **BayesMix** took around 10s to be executed.

At every MCMC iteration we obtain a sample from the posterior distribution of the random partition of the stations. In order to provide useful insights on the clusters, we need to specify a point estimate. Under the Bayesian paradigm, the canonical approach is to introduce a loss function and select as point estimate the minimiser of the posterior expectation of such loss function. In this work, we rely on the Variation of Information (VI) loss function [5]. To compute the optimal partition, we rely on the randomised greedy search algorithm in the **salso** R package [6].

In Fig. 1 we display the monitoring stations on the map and the time series dataset, coloured according to the posterior point estimate of the clustering with respect to the VI loss function. First of all, we observe that the inferred clusters have a clear spatial interpretation: the blue and green clusters groups together the monitoring stations that are in mountain locations, i.e., where the altitude is high; the yellow cluster collects the stations close to the Adriatic sea, in Emilia-Romagna, while the red cluster groups the stations in the hinterland. If we then colour the observed time series according to the estimated clusters, we see that our model can detect different trends in the concentrations of PM_{10} : for instance, the red cluster is associated to higher concentration levels of pollutant

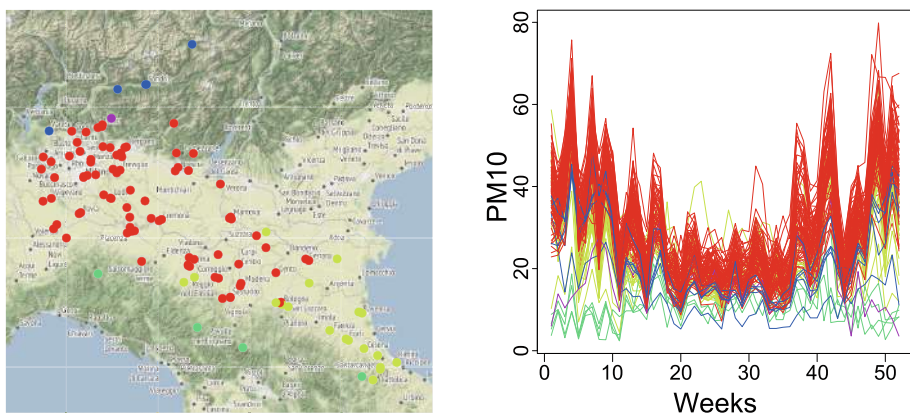


Fig. 1. Clustering of spatial time series: monitoring stations in the Po Valley map (left) and raw data (right) coloured according to the posterior point estimate of the cluster under VI loss function.

and exhibit a clear U-shape while the green and yellow clusters are associated to a lower and more constant concentration of fine dust throughout all year.

Acknowledgements. We thank Davide Carnevali, Davide Gurrieri, Sofia Moroni, Sara Rescalli, Eugenio Varetti and Chiara Zelioli, M. Sc. students in Mathematical Engineering at Politecnico di Milano, for the implementation of the **C++** code for **BayesMix** and the plots in Sect. 4. Both authors have been partially supported by MUR, Grant Dipartimento di Eccellenza 2023-2027. Alessandra Guglielmi has also been supported by MUR - PRIN 2022 - Grant no. 2022CLTYP4, funded by the European Union - Next Generation EU.

References

1. Page, G.L., Quintana, F.A.: Spatial product partition models. *Bayesian Anal.* **11**(1), 265–298 (2016)
2. Beraha, M., Guindani, B., Gianella, M., Guglielmi, A.: BayesMix: Bayesian mixture models in C++. *arXiv preprint [arXiv:2205.08144](https://arxiv.org/abs/2205.08144)* (2022)
3. Prado, R., West, M.: Time Series: Modeling, Computation, and Inference. Chapman and Hall/CRC (2010)
4. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* **9**(2), 249–265 (2000)
5. Meilă, M.: Comparing clusterings - an information based distance. *J. Multivar. Anal.* **98**(5), 873–895 (2007)
6. Dahl, D.B., Johnson, D.J., Müller, P.: Search algorithms and loss functions for Bayesian clustering. *J. Comput. Graph. Stat.* **31**(4), 1189–1201 (2022)



Statistical Methods for Evaluating the Environmental Sustainability

Andrea Gilardi^(✉) and Francesca Ieva

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
andrea.gilardi@polimi.it

Abstract. This paper represents a preliminary attempt at exploring a series of statistical methods for evaluating the environmental sustainability of urban and rural territories. We briefly present and discuss modern statistical techniques to study the relationships between a set of air pollutants. In particular, we focus on multivariate methods to explore air quality data after encoding the atmospheric measurements as covariance matrices that summarise the relationships among pollutants at different monitoring sites. In fact, a key property of covariance matrices is that they lie on a Riemannian manifold, and we exploit this fact to facilitate the exploratory analyses. Future directions of our work include extending the methods discussed here in a geostatistical setting, employing techniques such as Kriging for Riemannian data.

Keywords: air quality · geostatistical models · kriging · riemannian manifolds

1 Introduction

In recent years, air pollution is emerging as a pressing global and national concern, especially in some areas of the country near the Po Valley [4]. Recognizing this challenge, the European Commission recently stated that national and supranational policies should integrate climate change as a priority area to facilitate the transition towards a low-carbon economy. Therefore, during the last few years, the European Union (EU) promoted several projects targeting the environmental sector and, more precisely, the sustainability transition. These initiatives encompass diverse plans such as waste collection, development of renewable energy sources, or, as discussed here, air quality improvement.

Exploratory analyses, dimensionality reduction methods, and spatio-temporal air quality models are key to develop ex-ante forecasting or ex-post evaluation of the impact of the EU investments mentioned before, playing also a pivotal role to assess whether a project improved (or, at least, influenced) the environmental sustainability of a territory.

The starting point of the aforementioned techniques is typically a set of atmospheric data measured by a collection of monitoring stations spatially distributed

in a territory. These stations provide hourly or daily measurements for one or more air pollutants (e.g. NO₂, PM₁₀, O₃, ...), which are usually analysed one at a time. However, it's increasingly recognized that there exists a dynamic interplay between different chemical air pollutants in the atmosphere. Consequently, a growing body of literature advocates for a paradigm shift towards a “multipollutant approach to air quality” [1,6], acknowledging the intricate relationships between various pollutants and their combined impact on air quality.

For these reasons, the objective of this paper is to briefly discuss modern methods to jointly study a series of air pollutants over a region. We sketch a possible strategy to explore multi-site multi-pollutant atmospheric monitoring data using covariance matrices and exploiting the fact that they represent an importance instance of elements belonging to a Riemannian manifold (instead of classical Euclidean domain) [2]. The particular case study focuses on the analysis of PM₁₀, PM_{2.5}, and NO₂ levels in Lombardy during the time period 2021–2022.

The remaining parts of this article are structured as follows. The air pollution and the monitoring station data are described in Sect. 2, whereas Sect. 3 briefly introduces the main concepts underlying the analysis of Riemannian manifolds and summarises the future directions for our work.

2 Air Quality Monitoring Data

The point-level information regarding the multivariate air pollution measurements was derived from the *Air Quality e-Reporting database*¹ of the European Environmental Agency (EEA) using the corresponding download service. More precisely, the EEA database aggregates more than 12 years of daily and hourly time series measurements for a panel of air pollutants (e.g. NO₂, PM₁₀, O₃, ...) obtained from numerous monitoring stations scattered among the EU member states and several other cooperating countries (which are collectively named EEA38). The individual measurements and the meta-information regarding the monitoring stations involved in the observation process are uploaded into the EEA database by each single entity through a network named EIONet (Environmental Information and Observation Network). For the Italian case, this process is performed by an official national agency named ISPRA which collects and aggregates data for each region and autonomous province.

Considering the enormous amount of data stored in the complete database, the EEA distributes the raw time series using a binary size-efficient format named **parquet** that can be processed using a multi-language toolbox named *Apache Arrow*. After downloading the raw atmospheric measurements and matching them with the metadata of the monitoring grid, we filtered the observations recorded during the time period 2021–2022 pertaining to those stations that are located in the Lombardy region and jointly record PM₁₀, PM_{2.5}, and NO₂.

¹ The database can be browsed at the following link: <https://www.eea.europa.eu/en/datahub/datahubitem-view/3b390c9c-f321-490a-b25a-ae93b2ed80c1>. Data downloaded on March 2024.

The complete set of monitoring stations, (a sample of) the raw time series for the chosen pollutants, and the corresponding covariance matrices are reported in Fig. 1. As we can see, there are approximately 30 locations spread in different parts of the region where we jointly observe PM10, PM2.5, and NO2. To simplify the representation, we decided to focus on three of these thirty locations (which are displayed as red dots on the map), graphing the time series of pollutant levels recorded in these stations (central) and the corresponding variance-covariance matrices (right). There are varying degrees of correlation between the three pollutants (especially PM10 and PM2.5).

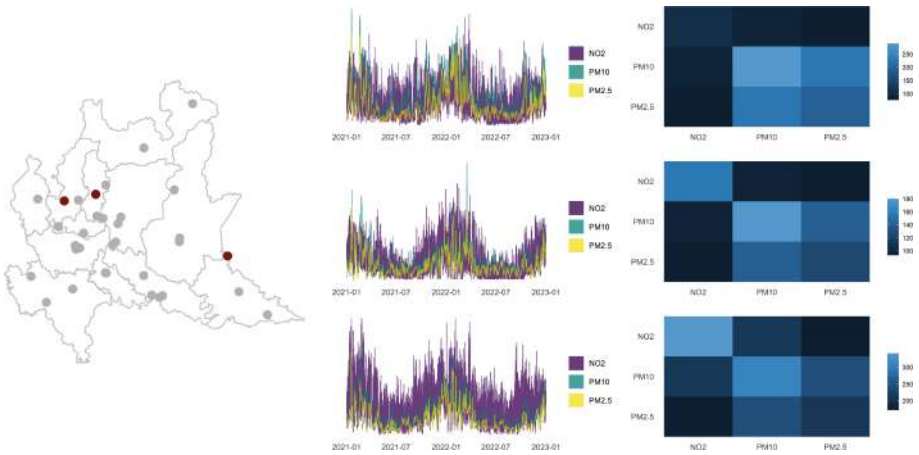


Fig. 1. (Left): Map of the monitoring stations in Lombardy that during the time period 2021–2022 jointly measured PM10, PM2.5, and NO2. Each dot represent a different station. (Center): Time series of the chosen pollutants for the three monitoring stations that are reported as red dots. (Right): Covariance matrices between the pollutants. Each plot corresponds to the adjacent time series.

3 Discussion and Future Works

Basics of Riemannian Manifolds

The objective of this section is to briefly sketch the main concepts underlying the analysis of data on Riemannian manifolds. In fact, as we already mentioned in the previous sections, symmetric positive-definite matrices (e.g. correlation and variance-covariance matrices) are objects lying on a Riemannian manifold, say $\mathcal{M}$, which is a smooth and differential manifold equipped with a *tangent space* at each point $p \in \mathcal{M}$. The tangent space is a vector space consisting of all possible tangent vectors to curves passing through p . Its properties let us compute the length of any curve passing through the manifold and, most importantly, find the *geodesic* (i.e. the curve of minimum length) between any two elements of $\mathcal{M}$.

Given a $p \times p$ symmetric positive definite matrix $\mathbf{M} \in \mathcal{M}$, we define its matrix logarithm as

$$\log(\mathbf{M}) = \mathbf{U} \text{diag}\{\log(\lambda_1), \dots, \log(\lambda_p)\} \mathbf{U}'$$

where $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{M}$ and $\lambda_1 > \dots > \lambda_p$ are its eigenvalues. The Frobenius inner product between two matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}$ is defined as

$$\langle \mathbf{M}_1, \mathbf{M}_2 \rangle_F = \text{tr}(\mathbf{M}_1' \mathbf{M}_2)$$

and the Frobenius norm is equal to $\|\mathbf{M}\|_F = \sqrt{\text{tr}(\mathbf{M}'\mathbf{M})} = \sqrt{\sum_{i=1}^n \lambda_i}$. These constructs let us compute the length of the geodesic between any two symmetric and positive definite covariance matrices:

$$d(\mathbf{M}_1, \mathbf{M}_2) = \|\log(\mathbf{M}_1) - \log(\mathbf{M}_2)\|_F.$$

The elements introduced in the previous paragraphs are crucial when modelling the dependence between spatial locations by means of (empirical) variogram models for sample covariance matrices. More precisely, given a set $\{s_i\}_{i=1}^n$ of n spatial location in a region D (e.g. the monitoring stations in Lombardy depicted in Fig. 1) for which we observed a sample covariance matrix $\mathbf{M}_i \in \mathcal{M}, i = 1, \dots, n$ (e.g. the covariance matrix among PM10, PM2.5 and NO2), the empirical variogram of the spatial process among the elements in the Riemannian manifold can be computed as

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{2|N(h)|} \sum_{(s_i, s_j) \in N(h)} d_g(\mathbf{P}_i, \mathbf{P}_j) \\ &= \frac{1}{2|N(h)|} \sum_{(s_i, s_j) \in N(h)} \|\log(\mathbf{P}_i) - \log(\mathbf{P}_j)\|_F \end{aligned} \quad (1)$$

where $N(h) = \{(s_i, s_j) \in D : h - \Delta h < \|s_i - s_j\| < h + \Delta h; i, j = 1, \dots, n\}$, Δh is a positive small quantity, $h > 0$, and $|N(h)|$ counts the number of couples (s_i, s_j) belonging to $N(h)$ [5]. The empirical variogram reported in Equation (1) represent a key step for the estimation of a Kriging model among spatial variance-covariance matrices that respect the underlying geometry of the manifold space.

Ongoing and Future Works

Clearly, the techniques described so far represent just the basics for a spatial analysis of air quality data lying on a Riemannian manifold, and we are currently working on testing such methods with the EEA data described in Sect. 2. Nevertheless, we believe they represent the right approach to tackle the challenging problem of correctly analysing the relationship between a set of air pollutants, taking into account the (geometrically corrected) spatio-temporal dynamics of these processes. In fact, a proper understanding of the (spatio-temporal) relationship and dynamics among pollutants can help decision makers shaping innovative and more effective policies targeting air quality improvement, improve

their spatial allocation, and help us better understand the impact of carbon abatement policies.

Finally, we point out that we aim to extend the techniques briefly described in the previous subsection considering, for example, the Kriging framework for manifold-valued random fields introduced in [5] and, possibly, extending it to a (separable or non-separable) spatio-temporal setting. In this case, we might consider estimating yearly variance-covariance matrices for each monitoring station and adapt the existing methods to the seasonal dimensions depicted in Fig. 1. Another interesting line of research raises when studying non-stationary random fields or irregularly shaped spatial domains (e.g. administrative regions with complex boundaries, holes, or barriers). In this case, following the work detailed in [3], we might consider the Random Domain Decomposition approach which will let us overcome the aforementioned problems using an ensemble of local models. Furthermore, the aforementioned technique is able to capture non-stationarities in the residual random field which cannot be explained by only modelling the drift term of the Kriging via a series of covariates (like altitude or population density). This aspect is particularly important when analysing multiple air pollutants since they typically exhibit complex non-linear relationship.

Acknowledgements. This study was funded by the European Union -Next GenerationEU, Mission 4, Component 2, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP D43C22003110001). 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.

The authors acknowledge the support by MUR, grant Dipartimento di Eccellenza 2023-2027.

References

1. Dominici, F., Peng, R.D., Barr, C.D., Bell, M.L.: Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology* **21**(2), 187–194 (2010)
2. Lee, J.M.: *Introduction to Riemannian Manifolds*, vol. 2. Springer, Cham (2018)
3. Menafoglio, A., Pigoli, D., Secchi, P.: Kriging riemannian data via random domain decompositions. *J. Comput. Graph. Stat.* **30**(3), 709–727 (2021)
4. Pernigotti, D., Georgieva, E., Thunis, P., Bessagnet, B.: Impact of meteorological modelling on air quality: summer and winter episodes in the Po valley (Northern Italy). *Int. J. Environ. Pollut.* **50**(1–4), 111–119 (2012)
5. Pigoli, D., Menafoglio, A., Secchi, P.: Kriging prediction for manifold-valued random fields. *J. Multivar. Anal.* **145**, 117–131 (2016)
6. Smith, A., Hua, J., de Foy, B., Schauer, J.J., Zavala, V.M.: Multi-site, multi-pollutant atmospheric data analysis using Riemannian geometry. *Sci. Total Environ.* **892**, 164064 (2023)



Hierarchical Clustering for Three-Way Data

Paolo Giordani¹, Susanna Levantesi¹, Andrea Nigri^{2(✉)}, and Donatella Vicari¹

¹ Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
{paolo.giordani,susanna.levantesi,donatella.vicari}@uniroma1.it

² Dipartimento di Economia, Management e Territorio, Università di Foggia, Foggia, Italy
andrea.nigri@unifg.it

Abstract. A novel clustering model is presented for three-way data that refer to a set of units on which variables are measured or collected at different occasions. The proposal originates from the CPclus model [9], where both clusters of units and components for variables and occasions are identified in a k -means based framework. Here we develop a hierarchical variant, called H-CPclus, which is implemented using a divisive approach, where the non-hierarchical model is applied recursively to obtain nested partitions. This allows the results to be displayed in a standard dendrogram fashion.

Keywords: Three-way data · dimensionality reduction · hierarchical clustering · Candecomp/Parafac

1 Introduction

Three-way data refer to a set of units on which some variables are collected in several occasions [4]. They can be stored in a three-way array or tensor, say $\mathbf{X}$ of order $(I \times J \times K)$, where I , J and K denote the numbers of units, variables and occasions, respectively. In other words, the available data can be seen as a collection of standard data matrices, one for each occasion. It is clear that the information is richer if compared with the standard (two-way) case, but, at the same time, poses new challenges because classical statistical methods fail to analyze three-way data properly. For this reason, in the next section we describe the most common tools for three-way data, namely, the so-called Candecomp/Parafac and Tucker3 models that represent three-way extensions of classical Principal Component Analysis. Although three-way data are usually summarized in terms of components, clustering for three-way data is receiving increasing attention. Clustering models for three-way data are briefly recalled in Sect. 3. As far as we know, most existing clustering models follow a non-hierarchical strategy. For this reason, a novel hierarchical clustering procedure, based on the Candecomp/Parafac model, will be proposed in Sect. 4. Some final remarks conclude the paper.

2 Three-Way Component Analysis

In this section we describe the Candecomp/Parafac [1,3] and the Tucker3 [8] models. For this purpose, it is convenient to convert the tensor $\underline{\mathbf{X}}$ into a matrix. This can be done by means of $\mathbf{X}_A$, a matrix of order $(I \times JK)$ obtained by collecting the data matrices pertaining to the K occasions next to each other. Letting $\mathbf{X}_k$ be the matrix of order $(I \times J)$ corresponding to occasion k ($k = 1, \dots, K$), we get

$$\mathbf{X}_A = [\mathbf{X}_1 \cdots \mathbf{X}_k \cdots \mathbf{X}_K]. \quad (1)$$

2.1 The Tucker3 Model

The Tucker3 model, hereinafter T3, can be formulated as

$$\mathbf{X}_A = \mathbf{A}\mathbf{H}_A(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E}_A, \quad (2)$$

where the symbol $\otimes$ denotes the Kronecker product. $\mathbf{A}$ ($I \times P$), $\mathbf{B}$ ($J \times Q$) and $\mathbf{C}$ ($K \times R$) are the component matrices for the units, variables and occasions, respectively. The T3 model assumes P , Q and R components for units, variables and occasions, respectively, allowing for different levels of complexity in the low-dimensional configurations of units, variables and occasions. The interactions between the three sets of components are explained by the core matrix $\mathbf{H}_A$, representing the matricization of the core tensor $\underline{\mathbf{H}}$. Finally $\mathbf{E}_A$ is the error term.

2.2 The Candecomp/Parafac Model

The Candecomp/Parafac model, briefly CP, looks for the same number S of components for units, variables and occasions. It can be expressed as

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_A, \quad (3)$$

where $\odot$ is the Khatri-Rao product (columnwise Kronecker product) and $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are again the component matrices for the units, variables and occasions, but now of size $(I \times S)$, $(J \times S)$ and $(K \times S)$, respectively.

T3 and CP are two different models satisfying different properties. However, the CP model can be seen as a special case of T3. This is easily observed by equivalently rewriting CP as

$$\mathbf{X}_A = \mathbf{A}\mathbf{I}_A(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E}_A, \quad (4)$$

where $\mathbf{I}_A$ is the matricization of the three-way identity array. By comparing (4) and (2), we can observe that CP is a special version of T3 with $P = Q = R = S$ and $\mathbf{H}_A = \mathbf{I}_A$.

3 Three-Way Clustering

The T3 and CP models represent two strategies for summarizing three-way data through components. However, it can be of interest to summarize the data by seeking clusters. Although there exist proposals for partitioning units, variables and occasions simultaneously (see, e.g., [6]), the most common case is to partition one of them, without loss of generality, the units. In such a case, the goal is to cluster the units and to reduce the dimensionality of variables and occasions by components. In this respect, the T3clus model [7] has been developed. In brief, T3clus is an extension of T3 where the component matrix $\mathbf{A}$ is replaced by the allocation matrix $\mathbf{U}$. Several generalizations of CP in a clustering framework have been proposed in the literature. Among others, we refer to [9–11] and references therein.

3.1 The CPclus Model

In [9], the so-called CPclus model is introduced. As for T3clus, the main idea is to consider an allocation matrix $\mathbf{U}$. We start from (3) and denote the number of clusters by G . The CPclus model can be written as

$$\mathbf{X}_A = \mathbf{U}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_A, \quad (5)$$

where $\mathbf{U}$, of order $(I \times G)$, contains elements equal to either 0 or 1 and is such that the sum of each row elements is equal to 1. Notice that $\mathbf{B}$ and $\mathbf{C}$ are still the component matrices for the variables and occasions, but now have sizes $(J \times G)$ and $(K \times G)$, respectively. This allows us to emphasize how CPclus behaves. The generic g -th cluster of the units is associated to the g -th column of $\mathbf{U}$, $\mathbf{B}$ and $\mathbf{C}$, denoted $\mathbf{u}_g$, $\mathbf{b}_g$ and $\mathbf{c}_g$, respectively. The elements of $\mathbf{u}_g$ equal to one detect the units belonging to cluster g . Such a cluster can be interpreted in terms of variables and occasions by looking at the elements of $\mathbf{b}_g$ and $\mathbf{c}_g$. Therefore, consistently with CP, we can observe a one-to-one relation between cluster g and component g ($g = 1, \dots, G$). In particular, if a unit is assigned to cluster g , then it is fitted by the g -th variable and occasion components. Obviously, such a peculiarity does not hold for Tucker3-based methods.

It follows that the interpretation of the T3clus solution is more complex than the CPclus one because the clusters for the units depend on all the variable and occasion components and their interactions given by the core $\mathbf{H}$.

Parameter estimation is carried out in the least-squares sense by minimizing the sum of squares of the error term. The CPclus solution is the minimizer of the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{B}, \mathbf{C}} \quad & f = \|\mathbf{E}_A\|^2 = \|\mathbf{X}_A - \mathbf{U}(\mathbf{C} \odot \mathbf{B})^T\|^2 \\ \text{s.t.} \quad & u_{ig} \in \{0, 1\}, i = 1, \dots, I, g = 1, \dots, G, \\ & \sum_{g=1}^G u_{ig} = 1, g = 1, \dots, G, \end{aligned} \quad (6)$$

being $\|\cdot\|$ the Frobenius norm of matrices. An Alternating Least Squares algorithm can be used for finding the minimum of problem (6). See, for further details, [9].

All of the above methods follow a non-hierarchical approach to clustering, as they adopt a k -means-based [5] strategy where centroids are expressed in terms of a low-dimensional configuration of variables and occasions. As such, T3clus and CPclus represent three-way generalizations of the reduced k -means [2].

4 The H-CPclus Procedure

In many practical situations, it can be useful to follow a hierarchical strategy because it can offer richer information and insight into the data. By increasing the number of clusters, hierarchical clustering methods provide a more refined description of the data at the cost of losing parsimony and increasing complexity. Obviously, such comments also hold in three-way clustering. For all of these reasons, we develop a hierarchical procedure for clustering three-way data. This is based on CPclus, taking into account its simplicity. We refer to such a new procedure as Hierarchical CPclus (H-CPclus).

As for any non-hierarchical clustering tool, by running CPclus with G and $G+1$ clusters, the partition with G clusters is not dependent on the partition with $G+1$ clusters. In H-CPclus, to better evaluate how model complexity varies with respect to G , we adopt a divisive approach by working in a top-down manner. At the beginning of the procedure, there is one cluster ($G = 1$) consisting of all I objects. In each of the following steps, one cluster is split into two clusters using CPclus with two clusters/components, and the procedure stops after $I - 1$ steps with $G = I$ singleton clusters. The cluster to be split is chosen to maximize the reduction of the contribution to the loss function. This consists in evaluating the contribution of each cluster to the loss function in (6) by taking into account that

$$f = \sum_{g=1}^G f_g, \quad (7)$$

with

$$f_g = \|\mathbf{X}_a^{(g)} - \mathbf{1}_{I_g}(\mathbf{c}_g \otimes \mathbf{b}_g)^T\|^2. \quad (8)$$

In (8), $\mathbf{X}_a^{(g)}$ is the submatrix of $\mathbf{X}_a$ obtained by considering only the rows pertaining to the units belonging to cluster g and $\mathbf{1}_{I_g}$ is the vector with all elements equal to 1 of length I_g , where I_g denotes the number of units assigned to cluster g .

H-CPclus can be summarized as follows.

Initialization

Apply CPclus with two clusters/components to the full data;

Repeat

Starting from the partition in G clusters, find the partition in $G + 1$ clusters by

applying CPclus with two clusters/components to the cluster g^* corresponding to the maximum f_{g^*} in (8);

Until

I singleton clusters are found.

The H-CPclus results can be displayed by using a standard dendrogram where the heights of the nodes are the (maximum) values of f_g used for splitting the clusters. The key-point of the proposal is that we can look at f_g as a sort of CP-based within-deviance where the centroids are based on the one-component CP decomposition. At each step, the procedure splits the cluster with the highest degree of heterogeneity, i.e. the cluster characterized by the maximum value of f_g , into two clusters.

5 Final Remarks

In this paper, we have addressed the clustering problem for three-way data. Focusing our attention on finding a partition of units, this is usually done by identifying groups for the units and components for the variables and the occasions. In other words, it consists of searching for clusters with centroids that have a low-dimensional configuration through the variable and occasion components. While most of the literature has dealt with the problem by adopting a non-hierarchical strategy, we have developed a hierarchical procedure, called H-CPclus. H-CPclus provides a hierarchy of partitions, offering a more refined description of the three-way data as the number of clusters increases. Such a hierarchy is obtained by a divisive approach, where the non-hierarchical counterpart of H-CPclus, i.e. CPclus, is applied recursively with two clusters.

During the conference, the results of a real-life application of H-CPclus will be presented in order to show its effectiveness.

References

1. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n -way generalization of Eckart-Young decomposition. *Psychometrika* **35**, 283–319 (1970)
2. De Soete, G., Carroll, J.D.: k -means clustering in a low-dimensional Euclidean space. In: *New Approaches in Classification and Data Analysis*, pp. 212–219. Springer, Heidelberg (1994)
3. Harshman, R.A.: Foundations of the Candecomp/Parafac procedure: models and conditions for an ‘explanatory’ multi-modal factor analysis. *UCLA Work. Pap. Phon.* **16**, 1–84 (1970)
4. Kroonenberg, P.M.: *Applied Multiway Data Analysis*. Wiley, Hoboken (2008)
5. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
6. Papalexakis, E.E., Sidiropoulos, N.D., Bro, R.: From K -means to higher-way co-clustering: multilinear decomposition with sparse latent factors. *IEEE Trans. Signal Process.* **61**, 493–506 (2013)

7. Rocci, R., Vichi, M.: Three-mode component analysis with crisp or fuzzy partition of units. *Psychometrika* **70**, 715–736 (2005)
8. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311 (1966)
9. Vicari, D., Giordani, P.: CPclus: Candecom/Parafac clustering model for three-way data. *J. Classif.* **40**, 432–465 (2023)
10. Wilderjans, T.F., Ceulemans, E.: Clusterwise Parafac to identify heterogeneity in three-way data. *Chemom. Intell. Lab. Syst.* **129**, 87–97 (2013)
11. Wilderjans, T.F., Cariou, V.: CLV3W: a clustering around latent variables approach to detect panel disagreement in three-way conventional sensory profiling data. *Food Qual. Prefer.* **47**, 45–53 (2016)



Insights into Regression-Based Cross-Temporal Forecast Reconciliation

Daniele Girolimetto^(✉) and Tommaso Di Fonzo

Department of Statistical Sciences, University of Padova, Via C. Battisti 241, 35121
Padova, Italy

daniele.girolimetto@unipd.it

Abstract. We address some open-issues related to the relationships between sequential, iterative and optimal combination cross-temporal forecast reconciliation approaches proposed by [2]. We discuss the conditions under which a sequential (either first-cross-sectional-then-temporal, or first-temporal-then-cross-sectional) approach is equivalent to a fully (i.e., cross-temporally) coherent iterative heuristic. We also show that, for specific patterns of the error covariance matrix of the regression model on which the optimal combination approach grounds, iterative reconciliation ‘converges’ to the optimal combination solution. The reduction of the computing effort is evaluated in the experiment of forecasting hourly photovoltaic power generation considered by [3].

Keywords: Cross-temporal forecast reconciliation · Sequential · iterative · optimal combination approaches

1 Introduction

As effective decision-making depends on the support of accurate and coherent forecasts, forecast reconciliation has become a pivotal post-forecasting process aimed to improve the accuracy predictions for linearly constrained multiple time series [7, 9]. The need for reconciliation arises when base forecasts, often generated independently by separate organizational units or using distinct models for each variable, fail to align with the constraints linking the target-observed series. Classical reconciliation methods, such as bottom-up and top-down, have been widely employed to address cross-sectionally incoherent forecasts. The temporal coherence aspect is a natural extension, calling for the reconciliation of forecasts generated at different time granularities, such as monthly, quarterly, or annually, to ensure consistency between them [1]. The cross-temporal framework puts together the cross-sectional and temporal dimensions to achieve fully coherent forecasts [2]. In this setting, the complexity and dimensions of the problem grow very quickly along with the requested computational time and memory. This issue is tackled by [8] and [10] using sequential procedures that alternate uni-dimensional reconciliation approaches. On the other hand, [2] propose an iterative heuristic, extending the results so far, and an optimal (in least squares

sense) approach, that simultaneously encompasses contemporaneous and temporal coherency. In this paper, we discuss the conditions under which a sequential approach, is equivalent to a fully (i.e., cross-temporally) coherent iterative heuristic. We also show that, for specific patterns of the error covariance matrix of the regression model on which the optimal combination approach grounds, iterative reconciliation ‘converges’ to the optimal combination reconciliation solution. The reduction of the computing effort is evaluated in the experiment of forecasting hourly photovoltaic power generation considered by [3].

2 Regression-Based Cross-Temporal Reconciliation

Let $\mathbf{X}$ be the $[n \times (k^* + m)]$ matrix containing the values to be forecasted of a linearly constrained n -variate time series for all the temporal aggregation orders $k \in \mathcal{K}$ (e.g., for a quarterly time series, $\mathcal{K} = \{4, 2, 1\}$, where $k = 4$ denotes an annual, $k = 2$ semi-annual, and $k = 1$ quarterly, the highest frequency variable) where $m = \max_k \mathcal{K}$ and $k^* = \sum_{k \in \mathcal{K} \setminus k_1} \frac{m}{k}$. This matrix can be written as¹:

$$\mathbf{X} = [\mathbf{X}^{[m]} \mathbf{X}^{[k_{p-1}]} \dots \mathbf{X}^{[k_2]} \mathbf{X}^{[1]}] = \begin{bmatrix} \mathbf{U}^{[m]} & \mathbf{U}^{[k_{p-1}]} & \mathbf{U}^{[k_2]} & \mathbf{U}^{[1]} \\ \mathbf{B}^{[m]} & \mathbf{B}^{[k_{p-1}]} & \mathbf{B}^{[k_2]} & \mathbf{B}^{[1]} \end{bmatrix},$$

where $\mathbf{X}^{[k]} = \begin{bmatrix} \mathbf{U}^{[k]} \\ \mathbf{B}^{[k]} \end{bmatrix}$, $k \in \mathcal{K}$, is the $\left(n \times \frac{m}{k}\right)$ matrix containing the target values of the n temporally aggregated series of order k . As it is usual in the standard hierarchical/grouped forecasting [7], the n variables are split into a $[n_a \times (k^* + m)]$ matrix $\mathbf{U}$ and a $[n_b \times (k^* + m)]$ matrix $\mathbf{B}$, consisting in the data of, respectively, n_a upper and n_b bottom variables in the cross-sectional hierarchy, with $n = n_a + n_b$. It is worth noting that in the cross-temporal framework, the ‘very bottom’ variables are the highest-frequency bottom time series in matrix $\mathbf{B}^{[1]}$. Now, denote $\mathbf{S}_{cs}$ and $\mathbf{S}_{te}$ the structural (summing) matrices in the cross-sectional and temporal frameworks, respectively, mapping the relevant bottom variables into their complete counterpart vector. The structural representations valid for all n variables and all temporal aggregation orders $k \in \mathcal{K}$, which take into account either cross-sectional or temporal coherence, are given by $\mathbf{X} = \mathbf{S}_{cs}\mathbf{B}$ and $\mathbf{X}' = \mathbf{S}_{te}(\mathbf{X}^{[1]})'$, respectively. Since $\mathbf{X}^{[1]} = \mathbf{S}_{cs}\mathbf{B}^{[1]}$, and $\mathbf{B}' = \mathbf{S}_{te}(\mathbf{B}^{[1]})'$, we can express in matrix form the cross-temporal structural representation as well: $\mathbf{X}' = \mathbf{S}_{te}(\mathbf{B}^{[1]})'\mathbf{S}_{cs}'$. In addition, it is convenient considering the vectorized forms of the structural representations above:

$$\mathbf{x} = \mathbf{K}_{cs}\mathbf{b}, \quad \mathbf{x} = \mathbf{K}_{te}\mathbf{x}^{[1]}, \quad \mathbf{x} = \mathbf{K}_{ct}\mathbf{b}^{[1]},$$

¹ This section is largely based on the notation and the results by [2, 3, 5, 6]. Many technical details, not presented here for reasons of space, will be made available in the extended version of the paper.

where $\mathbf{x} = \text{vec}(\mathbf{X}')$, $\mathbf{b} = \text{vec}(\mathbf{B}')$, $\mathbf{x}^{[1]} = \text{vec}(\mathbf{X}^{[1]'})$, and $\mathbf{b}^{[1]} = \text{vec}(\mathbf{B}^{[1]'})$, $\mathbf{K}_{ct}$ is the cross-temporal structural matrix, encompassing both contemporaneous and temporal constraints, and

$$\mathbf{K}_{cs} = (\mathbf{S}_{cs} \otimes \mathbf{I}_{(k^*+m)}), \quad \mathbf{K}_{te} = (\mathbf{I}_n \otimes \mathbf{S}_{te}), \quad \mathbf{K}_{ct} = (\mathbf{S}_{cs} \otimes \mathbf{S}_{te}) = \mathbf{K}_{cs}\mathbf{K}_{te}.$$

Table 1 summarizes the results shown so far, and expresses the reconciled forecasts through the classical formula (see [7]) $\tilde{\mathbf{x}}_r = \mathbf{K}_r \mathbf{G}_r \hat{\mathbf{x}}$, $r \in \{cs, te, ct\}$, with $\mathbf{G}_r = (\mathbf{K}_r' \boldsymbol{\Omega}^{-1} \mathbf{K}_r)^{-1} \mathbf{K}_r' \boldsymbol{\Omega}^{-1}$, where $\mathbf{M}_r = \mathbf{K}_r \mathbf{G}_r$ are oblique projection matrices [9]. More precisely, $\tilde{\mathbf{x}}_{cs}$ is the oblique projection in the space $\mathcal{S}_{cs}$ spanned by the columns of $\mathbf{K}_{cs}$, and $\tilde{\mathbf{x}}_{te}$ is the oblique projection in the space $\mathcal{S}_{te}$ spanned by the columns of $\mathbf{K}_{te}$. In addition, $\tilde{\mathbf{x}}_{ct}$ is the oblique projection in the space $\mathcal{S}_{ct}$, given by the intersection of $\mathcal{S}_{cs}$ and $\mathcal{S}_{te}$ (i.e., $\mathcal{S}_{ct} = \mathcal{S}_{cs} \cap \mathcal{S}_{te}$), spanned by the columns of $\mathbf{K}_{ct}$. These three approaches are optimal in least squares sense, but only $\tilde{\mathbf{x}}_{ct}$ is cross-temporally coherent.

Table 1. Matrix and vectorized forms of the structural representation of an n -variate linearly constrained time series considered at all temporal aggregation orders, in the cross-sectional (cs), temporal (te), and cross-temporal (ct) frameworks. The column ‘Rec. forecasts’ reports the formula of the reconciled forecasts coherent with only cross-sectional ($\tilde{\mathbf{x}}_{cs}$), only temporal ($\tilde{\mathbf{x}}_{te}$), and both ($\tilde{\mathbf{x}}_{ct}$) constraints, respectively.

Regression-based reconciliation: structural representation			Constraints	
Matrix form	Vector form	Rec. forecasts	cs	te
$\widehat{\mathbf{X}} = \mathbf{S}_{cs} \mathbf{B} + \mathbf{E}$	$\mathbf{x} = \mathbf{K}_{cs} \mathbf{b} + \boldsymbol{\eta}$	$\tilde{\mathbf{x}}_{cs} = \mathbf{K}_{cs} \mathbf{G}_{cs} \hat{\mathbf{x}}$	yes	no
$\widehat{\mathbf{X}}' = \mathbf{S}_{te} \left(\mathbf{X}^{[1]} \right)' + \mathbf{E}'$	$\mathbf{x} = \mathbf{K}_{te} \mathbf{x}^{[1]} + \boldsymbol{\eta}$	$\tilde{\mathbf{x}}_{te} = \mathbf{K}_{te} \mathbf{G}_{te} \hat{\mathbf{x}}$	no	yes
$\widehat{\mathbf{X}}' = \mathbf{S}_{te} \left(\mathbf{B}^{[1]} \right)' \mathbf{S}_{cs}' + \mathbf{E}$	$\mathbf{x} = \mathbf{K}_{ct} \mathbf{b}^{[1]} + \boldsymbol{\eta}$	$\tilde{\mathbf{x}}_{ct} = \mathbf{K}_{ct} \mathbf{G}_{ct} \hat{\mathbf{x}}$	yes	yes

The formulae to compute $\tilde{\mathbf{x}}_{cs}$ and $\tilde{\mathbf{x}}_{te}$ have been used by [10] in two heuristic sequential reconciliation approaches, respectively called *cross-sectional-then-temporal* and *temporal-then-cross-sectional*, given by:

$$\tilde{\mathbf{x}}_{cst} = \mathbf{M}_{te} \tilde{\mathbf{x}}_{cs} = \mathbf{M}_{te} \mathbf{M}_{cs} \hat{\mathbf{x}}, \quad \tilde{\mathbf{x}}_{tcs} = \mathbf{M}_{cs} \tilde{\mathbf{x}}_{te} = \mathbf{M}_{cs} \mathbf{M}_{te} \hat{\mathbf{x}}.$$

In general, $\tilde{\mathbf{x}}_{cst}$ ($\tilde{\mathbf{x}}_{tcs}$) is not cross-sectionally (temporally) coherent. Theorem 1 in Sect. 3 establishes sufficient conditions under which the equality $\tilde{\mathbf{x}}_{cst} = \tilde{\mathbf{x}}_{tcs} = \tilde{\mathbf{x}}_{seq}$ holds. Furthermore, it is shown that under the same conditions, $\tilde{\mathbf{x}}_{seq}$ is equal to an optimal combination cross-temporal forecast reconciliation solution with a specific error covariance matrix.

The iterative cross-temporal reconciliation approach proposed by [2] consists in alternating forecast reconciliation along one single dimension (cross-sectional or temporal) in a cyclic fashion, until a convergence criterion is met:

$$\tilde{\mathbf{x}}_{ite,cst} = \lim_{j \rightarrow \infty} (\mathbf{M}_{te} \mathbf{M}_{cs})^j \hat{\mathbf{x}}, \quad \tilde{\mathbf{x}}_{ite,tcs} = \lim_{j \rightarrow \infty} (\mathbf{M}_{cs} \mathbf{M}_{te})^j \hat{\mathbf{x}},$$

where, provided convergence was achieved in both cases, $\tilde{\mathbf{x}}_{ite,cst}$ and $\tilde{\mathbf{x}}_{ite,tcs}$ are cross-temporally coherent, whereas in general $\tilde{\mathbf{x}}_{ite,cst} \neq \tilde{\mathbf{x}}_{ite,tcs}$. Theorem 2 in Sect. 3 establishes sufficient conditions under which the equality $\tilde{\mathbf{x}}_{ite,cst} = \tilde{\mathbf{x}}_{ite,tcs} = \tilde{\mathbf{x}}_{ite}$ holds, and shows that, under the same conditions, $\tilde{\mathbf{x}}_{ite}$ converges in norm to an optimal combination forecast reconciliation solution with a specific error covariance matrix.

3 Two New Results

The first result (Theorem 1) shows that if the error covariance matrices used in either steps of an iterative reconciliation approach are constant across levels and time granularities, (i) fully reconciled forecasts are obtained in a single (two-step) iteration, (ii) the final result does not depend on the order of application of the uni-dimensional reconciliation approaches, and (iii) the solution is equivalent to that obtained through an optimal combination approach using a separable covariance matrix having a Kronecker product structure. Then, in Theorem 2 it is shown a sufficient condition under which an iterative heuristic converges to an optimal combination approach.

Theorem 1. *Let $\mathbf{W}_j^{[k]}$, $k \in \mathcal{K}$, $j = 1, \dots, \frac{m}{k}$, be the $(n \times n)$ cross-sectional hierarchy error covariance matrix, and $\mathbf{\Omega}_i$, $i = 1, \dots, n$, the $[(k^* + m) \times (k^* + m)]$ error covariance matrix for the i -th series temporal hierarchy. If $\mathbf{W}_j^{[k]} = \mathbf{W}$, $\forall k \in \mathcal{K}$ and $\forall j = 1, \dots, \frac{m}{k}$, and $\mathbf{\Omega}_i = \mathbf{\Omega}$, $\forall i = 1, \dots, n$, then:*

1. *the iterative heuristic aimed to obtain cross-temporal reconciled forecasts reduces to a single (two-step) iteration. Furthermore, $\tilde{\mathbf{X}}_{tcs} = \tilde{\mathbf{X}}_{cst} = \tilde{\mathbf{X}}_{seq}$, where $\tilde{\mathbf{X}}_{tcs}$ ($\tilde{\mathbf{X}}_{cst}$) is the $[n \times (k^* + m)]$ matrix of the temporal-then-cross-sectional (cross-sectional-then-temporal) reconciled forecasts;*
2. *denoting $\tilde{\mathbf{X}}_{oct}$ the optimal combination reconciled forecasts obtained by using the covariance matrix $\mathbf{\Omega}_{ct} = \mathbf{W} \otimes \mathbf{\Omega}$, it is $\tilde{\mathbf{X}}_{seq} = \tilde{\mathbf{X}}_{oct}$.*

Proof. 1. If $\mathbf{W}_j^{[k]} = \mathbf{W}$, $\forall k \in \mathcal{K}$ and $\forall j = 1, \dots, \frac{m}{k}$, and $\mathbf{\Omega}_i = \mathbf{\Omega}$, $\forall i = 1, \dots, n$, the cross-sectional and temporal reconciled forecasts are respectively given by $\tilde{\mathbf{X}}_{cs} = \mathbf{M}_{cs}^* \hat{\mathbf{X}}$, and $\tilde{\mathbf{X}}_{te} = \hat{\mathbf{X}} \mathbf{M}_{te}^{*'}$, where $\mathbf{M}_{cs}^*$ and $\mathbf{M}_{te}^*$ are the $(n \times n)$ and $[(k^* + m) \times (k^* + m)]$, respectively, oblique projection matrices

$$\mathbf{M}_{cs}^* = \mathbf{S}_{cs}' (\mathbf{S}_{cs}' \mathbf{W}^{-1} \mathbf{S}_{cs})^{-1} \mathbf{S}_{cs}' \mathbf{W}^{-1}, \quad \mathbf{M}_{te}^* = \mathbf{S}_{te} (\mathbf{S}_{te}' \mathbf{\Omega}^{-1} \mathbf{S}_{te})^{-1} \mathbf{S}_{te}' \mathbf{\Omega}^{-1}.$$

It easy to check that $\tilde{\mathbf{X}}_{cst} = \mathbf{M}_{cs}^* \hat{\mathbf{X}} \mathbf{M}_{te}^{*'} = \tilde{\mathbf{X}}_{tcs}$. In addition, both $\tilde{\mathbf{X}}_{cst}$ and $\tilde{\mathbf{X}}_{tcs}$ are cross-temporally coherent, then $\tilde{\mathbf{X}}_{tcs} = \tilde{\mathbf{X}}_{cst} = \tilde{\mathbf{X}}_{seq}$.

2. Recalling that $\mathbf{S}_{ct} = \mathbf{S}_{cs} \otimes \mathbf{S}_{te}$, after a few algebraic steps we find that the optimal combination cross-temporal reconciled forecasts may be expressed as:

$$\tilde{\mathbf{x}}_{oct} = \mathbf{S}_{ct} (\mathbf{S}_{ct}' \mathbf{\Omega}_{ct}^{-1} \mathbf{S}_{ct})^{-1} \mathbf{S}_{ct}' \mathbf{\Omega}_{ct}^{-1} \hat{\mathbf{x}} = (\mathbf{M}_{cs}^* \otimes \mathbf{M}_{te}^*) \hat{\mathbf{x}}.$$

On the other hand, $\tilde{\mathbf{x}}_{cs} = (\mathbf{M}_{cs}^* \otimes \mathbf{I}_{k^*+m}) \hat{\mathbf{x}}$, and $\tilde{\mathbf{x}}_{te} = (\mathbf{I}_n \otimes \mathbf{M}_{te}^*) \hat{\mathbf{x}}$, so it easy to check that $\tilde{\mathbf{x}}_{tcs} = \tilde{\mathbf{x}}_{cst} = (\mathbf{M}_{cs}^* \otimes \mathbf{M}_{te}^*) \hat{\mathbf{x}} = \tilde{\mathbf{x}}_{oct}$. $\square$

Theorem 2. Let $\mathbf{W}^{[k]}$, $k \in \mathcal{K}$, be the cross-sectional hierarchy error covariance matrix with $\mathbf{W}_{ct} = \text{diag}(\mathbf{W}^{[m]}, \dots, \mathbf{W}^{[1]})$, and $\mathbf{\Omega}_i$, $i = 1, \dots, n$, the i -th series temporal hierarchy error covariance matrix with $\mathbf{\Omega}_{ct} = \text{diag}(\mathbf{\Omega}_1, \dots, \mathbf{\Omega}_n)$. Denote with $\mathbf{P}$ the commutation matrix operating on matrices of dimension $[n \times (k^* + m)]$ (e.g., $\mathbf{X}$), such that $\mathbf{P} \text{vec}(\mathbf{X}) = \text{vec}(\mathbf{X}')$. If $\mathbf{W}_{ct} = \mathbf{P}' \mathbf{\Omega}_{ct} \mathbf{P}$, then the iterative reconciled forecasts converge in norm to the optimal combination reconciled forecasts with cross-temporal covariance matrix $\mathbf{\Omega}_{ct}$.

Proof. Let $\widetilde{\mathbf{X}}_{tcs}$ and $\widetilde{\mathbf{X}}_{cst}$ be the $[n \times (k^* + m)]$ matrix of the temporal-then-cross-sectional and cross-sectional-then-temporal iterative reconciled forecasts using $\mathbf{W}^{[k]}$ ($k \in \mathcal{K}$) and $\mathbf{\Omega}_i$ ($i = 1, \dots, n$) as the cross-sectional and temporal covariance matrices, respectively. The iterative solution obtained at a given complete recursion $J \geq 1$, may be written as the result of alternating oblique projections, such that $\widetilde{\mathbf{x}}_{tcs} = (\mathbf{M}_{cs} \mathbf{M}_{te})^J \widehat{\mathbf{x}}$, with

$$\begin{aligned} \mathbf{M}_{cs} &= \mathbf{K}_{cs} [\mathbf{K}'_{cs} (\mathbf{P} \mathbf{W} \mathbf{P}')^{-1} \mathbf{K}_{cs}]^{-1} \mathbf{K}'_{cs} (\mathbf{P} \mathbf{W} \mathbf{P}')^{-1} \\ \mathbf{M}_{te} &= \mathbf{K}_{te} [\mathbf{K}'_{te} \mathbf{\Omega}^{-1} \mathbf{K}_{te}]^{-1} \mathbf{K}'_{te} \mathbf{\Omega}^{-1} \end{aligned} \quad (1)$$

We want to prove that $\|\widetilde{\mathbf{x}}_{tcs} - \widetilde{\mathbf{x}}_{oct}\| = \|(\mathbf{M}_{cs} \mathbf{M}_{te})^J \widehat{\mathbf{x}} - \mathbf{M}_{ct} \widehat{\mathbf{x}}\| \xrightarrow{J \rightarrow +\infty} 0$. Since $\mathbf{\Omega}$ is p.d., we can express its inverse as $\mathbf{\Omega}^{-1} = \mathbf{Q}' \mathbf{Q}$, where $\mathbf{Q}$ is a (unique) upper triangular matrix with positive diagonal elements. Thus, denoting $\bar{\mathbf{x}} = \mathbf{Q} \widehat{\mathbf{x}}$, it can be shown that $\widetilde{\mathbf{x}}_{cs} = \mathbf{Q}^{-1} \overline{\mathbf{M}}_{cs} \bar{\mathbf{x}}$, $\widetilde{\mathbf{x}}_{te} = \mathbf{Q}^{-1} \overline{\mathbf{M}}_{te} \bar{\mathbf{x}}$, $\widetilde{\mathbf{x}}_{oct} = \mathbf{Q}^{-1} \overline{\mathbf{M}}_{ct} \bar{\mathbf{x}}$, where $\overline{\mathbf{M}}_r = \overline{\mathbf{K}}_r (\overline{\mathbf{K}}'_r \overline{\mathbf{K}}_r)^{-1} \overline{\mathbf{K}}'_r$, with $\overline{\mathbf{K}}_r = \mathbf{Q} \mathbf{K}_r$, $r \in \{cs, te\}$, and $\overline{\mathbf{M}}_{ct} = \overline{\mathbf{S}}_{ct} (\overline{\mathbf{S}}'_{ct} \overline{\mathbf{S}}_{ct})^{-1} \overline{\mathbf{S}}'_{ct}$, with $\overline{\mathbf{S}}_{ct} = \mathbf{Q} \mathbf{S}_{ct}$, are orthogonal projection matrices onto the linear sub-spaces spanned by the columns of, respectively, $\mathbf{Q} \mathbf{S}_{ct}$, $\mathbf{Q} \mathbf{K}_{cs}$, and $\mathbf{Q} \mathbf{K}_{te}$. The oblique projection matrices (1) can thus be written using the orthogonal projection matrices $\overline{\mathbf{M}}_r = \overline{\mathbf{K}}_r (\overline{\mathbf{K}}'_r \overline{\mathbf{K}}_r)^{-1} \overline{\mathbf{K}}'_r$, $r \in \{cs, te\}$, such that $\mathbf{M}_r = \mathbf{Q}^{-1} \overline{\mathbf{M}}_r$. Then, $(\mathbf{M}_{cs} \mathbf{M}_{te})^J \widehat{\mathbf{x}} = \mathbf{Q}^{-1} (\overline{\mathbf{M}}_{cs} \overline{\mathbf{M}}_{te})^J \bar{\mathbf{x}}$ and thus,

$$\widetilde{\mathbf{x}}_{tcs} - \widetilde{\mathbf{x}}_{oct} = \mathbf{Q}^{-1} [(\overline{\mathbf{M}}_{cs} \overline{\mathbf{M}}_{te})^J - \overline{\mathbf{M}}_{ct}] \mathbf{Q} \widehat{\mathbf{x}}.$$

As the alternating orthogonal projections onto two subspaces converges in norm to the projection onto the intersection of the two subspaces², we have $\|\widetilde{\mathbf{x}}_{tcs} - \widetilde{\mathbf{x}}_{oct}\| \xrightarrow{J \rightarrow +\infty} 0$. The convergence of $\widetilde{\mathbf{x}}_{cst}$ to $\widetilde{\mathbf{x}}_{oct}$ is easily obtained by inverting the order in which matrices $\mathbf{M}_{cs}$ and $\mathbf{M}_{te}$ are applied in the previous proof. $\square$

4 Reduction in the Computational Effort

To evaluate the computational effectiveness of Theorems 1 and 2, we compare the iterative (tcs) and the optimal approaches for the forecasting experiment by [3]. This application involves 318 simulated photovoltaic plants in California, whose hourly ($m = 24$) irradiation data are organized in three levels for a total of 324 variables. The reconciliation approaches considered are:

² Thanks to Prof. Tommaso Proietti for suggesting considering alternating projections.

1. $ite(ols)$ with $\mathbf{W} = \mathbf{I}_n$ and $\mathbf{\Omega} = \mathbf{I}_{(k^*+m)}$ equivalent to $oct(ols)$ with $\mathbf{\Omega}_{ct} = \mathbf{I}_{n(k^*+m)}$;
2. $ite(struc)$ with $\mathbf{W} = \text{diag}(\mathbf{S}_{cs}\mathbf{1}_{n_b})$ and $\mathbf{\Omega} = \text{diag}(\mathbf{S}_{te}\mathbf{1}_m)$ equivalent to $\mathbf{\Omega}_{ct} = \mathbf{I}_{n(k^*+m)}$ and $oct(struc)$ with $\mathbf{\Omega}_{ct} = \text{diag}(\mathbf{S}_{ct}\mathbf{1}_{mn_b})$;
3. $ite(wlsv)$, iterative heuristic alternating the temporal series variance scaling and the cross-sectional series variance approaches, converges to $oct(wlsv)$, the cross-temporal series variance scaling approach³.

The equivalences in points 1 and 2 are provided by Theorem 1, and the convergence in point 3 by Theorem 2. Figure 1 shows the computational times of iterative and optimal combination approaches for the 350 replications of the forecasting experiment. In all cases, the optimal approach requires shorter computational times, almost always below two seconds, and this difference appears pronounced for $ite(wlsv)$, where the number of iterations exceeds 1.

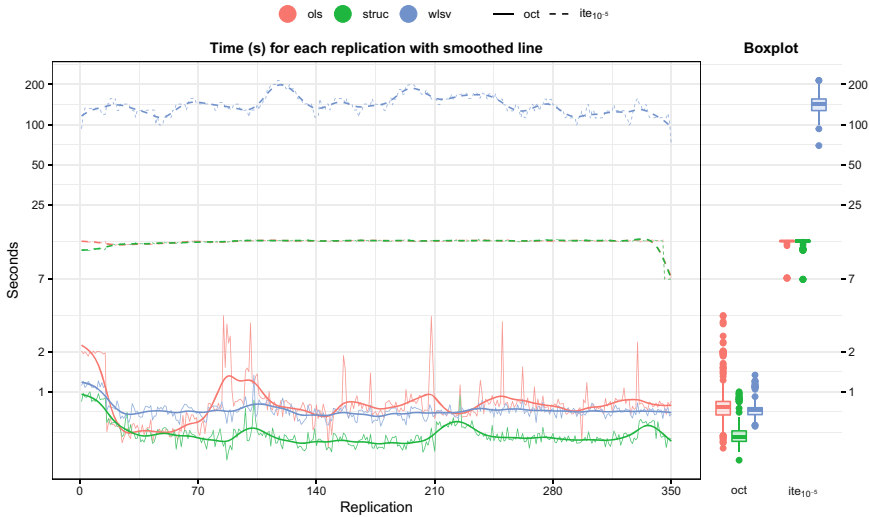


Fig. 1. Computational times of iterative and optimal reconciliation approaches (*ols*, *struc*, and *wlsv*). System's hardware and software specifications: Intel(R) Core(TM) i7-10700 CPU 2.90GHz 2.90 GHz, 64 GB RAM, R v4.1.2 and FoReco v0.2.5 [4].

References

1. Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., Petropoulos, F.: Forecasting with temporal hierarchies. *Eur. J. Oper. Res.* **262**(1), 60–74 (2017)
2. Di Fonzo, T., Girolimetto, D.: Cross-temporal forecast reconciliation: optimal combination method and heuristic alternatives. *Int. J. Forecast.* **39**(1), 39–57 (2023)

³ See [2] for more details about the different approaches.

3. Di Fonzo, T., Girolimetto, D.: Spatio-temporal reconciliation of solar forecasts. *Sol. Energy* **251**, 13–29 (2023)
4. Girolimetto, D., Di Fonzo, T.: FoReco: Forecast reconciliation (2023). <https://danigiro.github.io/FoReco/>
5. Girolimetto, D., Athanasopoulos, G., Di Fonzo, T., Hyndman, R.J.: Cross-temporal probabilistic forecast reconciliation: methodological and practical issues. *Int. J. Forecast.* (2023)
6. Girolimetto, D., Di Fonzo, T.: Point and probabilistic forecast reconciliation for general linearly constrained multiple time series. *Stat. Methods Appl.* (2023)
7. Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L.: Optimal combination forecasts for hierarchical time series. *Comput. Stat. Data Anal.* **55**(9), 2579–2589 (2011)
8. Kourentzes, N., Athanasopoulos, G.: Cross-temporal coherent forecasts for Australian tourism. *Ann. Tour. Res.* **75**, 393–409 (2019)
9. Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., Hyndman, R.J.: Forecast reconciliation: a geometric view with new insights on bias correction. *Int. J. Forecast.* **37**(1), 343–359 (2021)
10. Yagli, G.M., Yang, D., Srinivasan, D.: Reconciling solar forecasts: sequential reconciliation. *Sol. Energy* **179**, 391–397 (2019)



Explainable Decision Tree Ensembles

Agostino Gnasso^{1(✉)}, Massimo Aria¹, Carmela Iorio¹, and Marjolein Fokkema²

¹ Department of Economics and Statistics, University of Naples Federico II,
Naples, Italy

agostino.gnasso@unina.it

² Department of Methods and Statistics, University of Leiden,
Leiden, The Netherlands

Abstract. Today, there is a growing demand for models that not only predict outcomes, but also explain the decision-making processes involved. This is particularly true in domains where decisions have significant consequences, such as in the financial and health care sectors. Whether diagnosing patients, approving loans, or managing public health initiatives, decisions influenced by machine learning methods models can have profound implications. To address this explainability issue, we show the predictive power of ensemble decision trees with the transparent logic of single trees, providing deep understanding of ensemble reasoning.

Keywords: Machine Learning · Ensemble Methods · Explainable Ensemble Trees

1 Introduction

Machine learning (ML) methods have become a cornerstone of many areas of research and application, including medicine, business, finance, and health care. It is revolutionizing decision-making. The ability to manage large data sets, identify patterns, and make predictions with high accuracy and efficiency is what makes ML so attractive. But ML is not without its challenges. The most pressing is the inherent complexity of the algorithms, which often leads to the creation of ‘black box’ models that are powerful but not transparent. The need for interpretable and explainable machine learning is underlined by the growing role of these approaches in critical decision-making processes. As a result, it can be difficult for users to understand how decisions are being made. This affects trust and adoption in several areas. This opacity presents both technical and practical challenges. The concepts of interpretability and explainability are the answer to this problem. Although often used interchangeably in the literature, they encapsulate different aspects of the problem (see for example Aria et al., 2021; Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Lipton, 2018; Gaurav and Tiwari, 2023). Interpretability refers to the extent to which a human can understand cause and effect within an ML system. Explainability, on the other hand, goes inside the ML model to reveal the processes and logic underlying

its functioning. Together, the concepts of interpretability and explainability are key to understanding ML models, making them more accessible and reliable to both experts and non-experts. However, distinguishing between interpretability and explainability is crucial. Indeed, an understanding of the decisions of an ML model (interpretability) does not necessarily imply an understanding of its internal workings (explainability), and vice versa. This realisation has led to a growing body of research aimed at improving both aspects of ML models (Lisboa, 2013; Ribeiro et al., 2016; Guidotti et al., 2018).

Here, we develop an explanation tool for decision-tree ensembles. Based on the dissimilarity of observations in a forest, we derive a single tree that is easy to explain, while revealing the core associations underlying the predictions of the forest.

2 The Key Idea

Ensemble decision tree models, such as Random Forests and XGboost, are renowned for their higher performance across a multitude of prediction tasks. However, their inherent complexity and the aggregation of multiple decision trees often result in a lack of transparency, making it difficult for decision makers to understand the relationship between response variables and predictors and therefore fail to understand how a given prediction is made.

Aria et al. (2024) introduced a novel approach to enhancing the explainability of ensemble decision tree models without compromising their predictive accuracy. The proposed algorithm is based on the principle of simplification without loss of essential information. It achieves this by building a single tree structure that encapsulates the decision-making process of the model. The algorithm works on the dissimilarity matrix, constructed using the output of the ensemble decision tree model. The dissimilarity matrix summarizes the tree ensemble in a new way. Let D be the dissimilarity matrix used to evaluate the degree of discordance between pairs of observations within the decision tree ensemble. To obtain D , one computes the proportion of times two observations u_i and u_j co-occur in a terminal node among all trees in the ensemble. We define D as follows:

$$D = \begin{matrix} & \begin{matrix} u_1 & u_2 & u_3 & & u_j & & u_n \end{matrix} \\ \begin{matrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{matrix} & \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1j} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2j} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3j} & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & d_{i3} & \cdots & d_{ij} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & d_{nj} & \cdots & 0 \end{bmatrix} \end{matrix} = 1 - \begin{bmatrix} O_{ij} \end{bmatrix}$$

To compute D , one defines O_{ij} as a weighted co-occurrence measure, where the weight is a local goodness of fit of the terminal nodes in the weak learners

when they exist. Moreover, to maximize the difference in discordance between the parent and children nodes, one also defines the impurity measure. This structure is designed to be as simple as possible, minimizing complexity and preserving the accuracy of the original model. Once the dissimilarity matrix is constructed, the Explainable Ensemble Tree (E2Tree) algorithm constructs a new, decision tree structure. E2Tree algorithm generate a single tree that aims to minimize D in the terminal nodes. Each node in E2Tree represents a decision point that was essential throughout the ensemble, and the paths through the tree encapsulate the ensemble's collective decision making. The main advantage of the E2Tree algorithm lies in its ability to provide a clear and intuitive explanation of the ensemble model predictions. By consolidating ensemble decision making into a single tree structure, E2Tree allows experts and non-experts to trace the logic behind predictions, increasing confidence in using that model and simplifying model insights. Furthermore, the E2Tree approach is not limited to a specific type of ensemble decision tree model. Its methodology is applicable to any decision tree-based ensemble model, making it a versatile tool for improving model explainability in various applications. Whether applied to financial risk assessment, medical diagnosis, or predicting customer behavior, E2Tree offers a way to balance the need for predictive accuracy with the growing demand for transparency in machine learning models.

Further details on the E2Tree algorithm are provided in Aria et al. (2024).

3 Conclusion

The Explainable Ensemble Tree approach represents an innovative advancement in the field of ensemble machine learning methodologies, especially those that incorporate decision trees as weak learners. This innovative approach is unique for the integration of a novel dissimilarity matrix combined with a dendrogram-like structure, which collectively plays a critical role in improving the explainability of complex models. A remarkable feature of the E2Tree methodology is its ability to address the often challenging question of explainability within complex machine learning algorithms, without sacrificing the predictive accuracy that is crucial to the effectiveness of these models. The importance of the E2Tree approach is further underlined by the results of rigorous experiments conducted on real datasets from the UCI Machine Learning Repository and simulated datasets. These experiments are able to validate the effectiveness of the E2Tree method, demonstrating its ability to not only preserve the high accuracy typically associated with ensemble methods but also to provide comprehensive graphical representation. This representation clarifies the complex relationships and interactions that exist between the different variables involved in the model. Such clarity in visualization and understanding is particularly advantageous in domains where decision-making processes are highly dependent on both the high predictive performance of models and the ability to clearly understand how these models reach their conclusions. The future aim is to extend E2Tree to cases that involve multivariate response variables. This could significantly broaden

the impact of the E2Tree method in the field of machine learning, opening new opportunities for its application and improving its usefulness in tackling more complex and varied data analysis challenges.

Acknowledgement. This research has been financed by the following research projects: PRIN-2022 “SCIK-HEALTH” (Project Code: 2022825Y5E; CUP: E53D2300-611006); PRIN-2022 PNRR “The value of scientific production for patient care in Academic Health Science Centres” (Project Code: P2022RF38Y; CUP: E53D23016650001)

References

- Aria, M., Cuccurullo, C., Gnasso, A.: A comparison among interpretative proposals for random forests. *Mach. Learn. Appl.* **6**(100), 094 (2021)
- Aria, M., Gnasso, A., Iorio, C., Pandolfo, G.: Explainable ensemble trees. *Comput. Stat.* **39**, 3–19 (2024)
- Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint [arXiv:1702.08608](https://arxiv.org/abs/1702.08608) (2017)
- Gaurav, D., Tiwari, S.: Interpretability vs explainability: the black box of machine learning. In: 2023 International Conference on Computer Science, pp. 523–528. IEEE, Information Technology and Engineering (ICCoSITE) (2023)
- Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pp. 80–89. IEEE (2018)
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **51**(5), 1–42 (2018)
- Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* **16**(3), 31–57 (2018)
- Lisboa, P.J.: Interpretability in machine learning—principles and practice. In: International Workshop on Fuzzy Logic and Applications, pp 15–21. Springer, Cham (2013)
- Ribeiro, M.T., Singh, S., Guestrin, C.: “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 (2016)



The Study of the Empowerment in Two African Countries Using the Bayesian Networks

Silvia Golia^(✉)

University of Brescia, Cda. S. Chiara 50, 2222 Brescia, Italy

silvia.golia@unibs.it

Abstract. Empowerment is a complex and challenging construct to define and measure and, following the Kabber's approach, it encompasses three interconnected dimensions: resources, agency, and achievements. In essence, women's empowerment focuses on expanding choices, dismantling barriers, and challenging unequal power dynamics within societies to promote greater agency and decision-making for women. In this study we want to investigate the relationship between three measures of the empowerment (educational level, decision-making mainly in the domestic sphere and attitudes towards violence) and some other characteristic of the socio-economical context in which the target women, that is married women, live. The aim is achieved by using the Bayesian networks (BNs). In the analysis, two different African countries are taken into account, that is Congo Democratic Republic, which is a poor country, underdeveloped in the demographic transition, and Egypt, which is a relatively well-off country, ahead with the demographic transition. The interest is in investigating the form of the relations within the single country and between the two countries and the BNs were found useful for both purposes.

Keywords: Empowerment · Bayesian Networks · Kabber's approach

1 Introduction

Empowerment is a complex and challenging construct to define and measure. Despite this difficulty, there have been numerous efforts to conceptualize it. In the literature, terms like autonomy in decision-making or gender equality are often used interchangeably with empowerment. However, these concepts represent static elements within the broader dynamic process of empowerment, which involves the progressive increase of autonomy and equality [2]. Empowerment is not a fixed state but rather a continuous journey [7], where individuals gain the ability to make choices that were previously denied to them [2]. As individuals become empowered, they also acquire autonomy in decision-making, and their choices lead to specific and meaningful outcomes in their lives [1]. Empowerment encompasses three interconnected dimensions: resources, agency, and achievements, as proposed by Kabeer [2]. Resources are essential but not enough on their own to provide choices. Examples of resources include education, economic means, and financial literacy. The second dimension, agency, refers to the power individuals have to

control their life choices and set their own goals, which leads to decisions based on a sense of purpose, motivation, or meaning. Personal awareness, or a sense of agency, is crucial for acting freely [7] and having the ability to make choices. Alongside resources and agency, the third element of empowerment is measurable achievements, which represent the actual outcomes resulting from the choices made. For women to experience true decision-making power, they must first perceive themselves as capable and deserving of making choices [1]. This requires breaking down negative social perceptions, building capacity through resources and agency, and having the right to participate and influence decisions in their own lives. In essence, women's empowerment focuses on expanding choices, dismantling barriers, and challenging unequal power dynamics within societies to promote greater agency and decision-making for women. If a woman is denied the right to do something, regardless of her reasons, it indicates a lack of empowerment in one of the dimensions described by Kabeer [2, 3]. In this study we want to investigate the relationship between three aspects of the empowerment, that is educational level, decision-making mainly in the domestic sphere, and attitudes towards violence, and some other characteristic of the socio-economical context in which married women live. The aim is achieved by using the Bayesian networks.

2 Bayesian Networks

Probabilistic networks are graphical models that explicit through a graph, the interactions among a set of variables represented as vertices or nodes of the graph. Any pair of unconnected nodes of the graph indicates (conditional) independence between the variables represented by these nodes under certain conditions that can be read from the graph itself. Hence, a probabilistic network captures a set of (conditional) dependence and independence properties associated with the variables represented in the network [4]. BN belong to the class of probabilistic networks [6]. The underlined graph is called Directed Acyclic Graphs (DAG). A DAG G is composed by the pair $G = (\mathbf{V}, E)$, where $\mathbf{V}$ is a finite set of distinct nodes, $\mathbf{V} = \{V_i\}_{i=1}^k$, which correspond to a set of random variables $\mathbf{X} = \{X_{V_i}\}_{i=1}^k$ indexed by $\mathbf{V}$, and $E \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed edges between pairs of nodes in $\mathbf{V}$. An ordered pair $(V_i, V_j) \in E$ denotes a directed edge ($\rightarrow$) from node V_i to node V_j ; V_i is said to be a parent of V_j whereas V_j a child of V_i . The set of parents of a node V shall be denoted by $pa(V)$. A BN over the variables in $\mathbf{X}$ is defined as the triplet $(\mathbf{X}, \mathbf{G}, \mathbf{P})$, where $\mathbf{G}$ is a DAG and $\mathbf{P}$ is a set of conditional probability distributions containing one distribution of the $P(X_v|X_{pa(v)})$ kind for each $X_v \in \mathbf{X}$, where $X_{pa(v)}$ indicates the set of parent variables of X_v . The joint probability distribution of $\mathbf{X}$ is factorized as follows:

$$P(\mathbf{X}) = \prod_{v \in \mathbf{V}} P(X_v|X_{pa(v)})$$

So, a BN can be described in terms of a qualitative component, that is the DAG, and a quantitative component, consisting of the joint probability distribution.

It is important to underline that, although the arrows in a BN are directed, they do not have a causal interpretation unless they belong to a causal BN, and information can flow both ways according to strict rules. For each node in a BN, it is possible to identify its *Markov blanket* (MB), which is the set of nodes consisting of its parents, its children, and any other parents of its children. An interesting property of the MB of a node consists in rendering the node conditionally independent of every other node in the graph given the MB. So, this implies that the MB of the node V_j identifies the set of strongly relevant variables for X_{V_j} , that is the set of variables that carry information about X_{V_j} that cannot be obtained from any other variable [5].

The construction of a BN follows two steps: the identification of the DAG, and the estimation of the conditional probability distributions $P(X_v | X_{pa(v)})$. The DAG can be obtained either manually, thanks to the knowledge of experts in the field, or automatically from data, applying an algorithm belonging to one of the following three broad categories: constraint-based, score-based, and hybrid algorithms. The method used in this paper to derive the DAG is the Hill Climbing (HC), which belongs to the class of the score-based algorithms and it consists in exploring the search space starting from an empty DAG and adding, deleting or reversing one arc at a time until the score considered can no longer be improved. Moreover, in order to evaluate the strength of the arcs which are present in the DAG learned on the data, we extracted 1000 bootstrap samples from the original data and on each of them we learned the corresponding network structure. The estimated strength of each arc was computed as the empirical frequency of each arc over the set of 1000 DAG learned from the bootstrap samples [9].

3 Data and Results

The data come from the Demographic Health Surveys, which are nationally representative surveys carried out in developing countries that gather evidence mainly in the domains of demography and health, thus enabling cross-country comparisons. We have considered two African countries, Congo Democratic Republic (CDR) and Egypt. The CDR is an extremely poor country and underdeveloped in the demographic transition, whereas the Egypt is a relatively well-off country, ahead with the demographic transition. Moreover, in terms of Human Development Index (HDI) Egypt owns a higher HDI (0.731) than CDR (0.479). So, it is of interest to study the relations between the variables considered in this study identified by the BN at the country level and also compare the resulting two BNs. The target statistical unit was married woman and the resulting sample sizes are 11380 (CDR) and 19412 (Egypt).

The variables considered in this study are the following:

- *Age* reports the woman's age factorized into 4 levels (0: >18 ; 1: 18–22; 2: 23–30; 4: >30).
- *Urban* indicates whether the person's de facto residence was in an urban or rural location; it is a factor variable with 2 levels.
- *Educllev* reports, for women who ever went to school, the highest level of school they attended; it is a factor variable with 4 levels (No education, Primary, Secondary, Higher).

- *AgeFrsMar* reports the woman's age at the start of her first marriage or union factorized into 4 levels (0: >18; 1: 18–22; 2: 23–30; 4: >30).
- *Wealth* refers to the relative wealth of the household where the woman lives; it is a factor variable with 5 levels (Poorest, Poorer, Middle, Richer, Richest).
- *Work* is computed as the combination of *CURRWORK*, which indicates whether the woman is currently working at a job or business, and *WKEARNTYPE*, which reports the type of earnings for the respondent's work. It is a factor variable with 3 levels (No current work, Work and paid, Work, but not paid).
- *IdealKid* reports the ideal total number of children the woman would have liked to have in her whole life factorized into 2 levels (0 if the ideal total number of kids of the woman is lower than the median of the ideal total number of kids of the country, 1 otherwise)
- *Cheb* reports the total number of children ever born to the respondent factorized into 2 levels (0 if the total number of children ever born of the woman is lower than the median of the total number of children ever born of the country, 1 otherwise)
- *Unmet* reports a woman's need for family planning; it is a factor variable with 3 levels (No unmet need FP, Not at risk, Unmet need)
- *DecBigHousPurch* indicates who, in the woman's household, usually makes decisions regarding large household purchases; it is a factor variable with 4 levels (Husband, Other, With someone else, Woman alone)
- *GoOut* reports if the husband is justified to beat his wife if she goes out without telling him; it is a factor variable with 2 levels (Yes, No). The response "Don't know" is treated as "Yes".

Figure 1 shows the BNs of the two countries. The analysis was performed using the *R* package *bnlearn* provided by Scutari [8]. In order to have a visualization of the strength of the arcs in the DAGs, strength derived from the bootstrap procedure described previously, the arcs were represented by arrows with increasing thickness in proportion to their strength, and by dashed lines for arcs with frequency less than 0.85.

Table 1 reports the list of arcs in the BNs with strength less than 0.85.

The two BNs show different structures which contain some common substructures, for example the one involving *Urban*, *Wealth* and *EducLev*. The percentage of arcs which appear in both DAGs with the same direction is 64.70% for CDR and 57.89% for Egypt, if we consider all the arcs regardless their robustness.

It is of interest to comment the role of three variables, *GoOut*, *DecBigHousPurch* and *EducLev* because they can be seen as three possible proxies for female empowerment. In fact, asking if it is justified that the husband beats his wife if she goes out without telling him can be seen as a proxy of women's empowerment and safety within the household. Knowing who, in woman's household, usually makes decisions regarding large household purchases refers to women's participation in household decision making. Finally, the level of education is often associated with greater social and economic opportunities for women.

There is a direct connection between the variables *GoOut* and *DecBigHousPurch* for both countries and this can be interpreted as an interdependency between the twos. Nevertheless, their MBs are different in the two countries, underlying their difference in terms of level of development and demographic transition. The MB for *GoOut* is

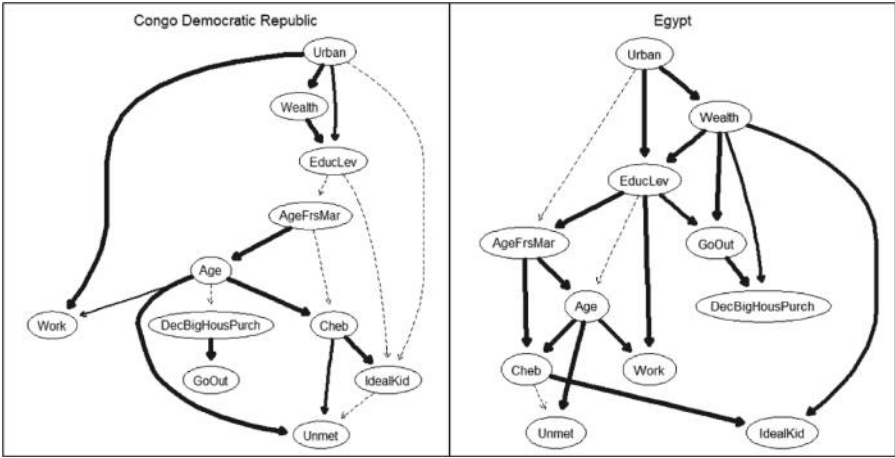


Fig. 1. BN for CDR and Egypt

Table 1. Arcs with strength less than 0.85.

CDR		Egypt	
Arcs	Strength	Arcs	Strength
Age → DecBigHousPurch	0.584	Urban → AgeFrsMar	0.617
Urban → IdealKid	0.793	EducLev → Age	0.517
EducLev → AgeFrsMar	0.716	Cheb → Unmet	0.752
EducLev → IdealKid	0.733		
IdealKid → Unmet	0.765		
AgeFrsMar → Cheb	0.737		

composed by *DecBigHousPurch* for CDR, whereas it also includes *EducLev* and *Wealth* for Egypt. So, in Egypt the educational level of the woman, the wealth of the household in which the woman lives and the person who usually makes decisions regarding large household purchases are sufficient to predict the attitude towards domestic violence whereas for CDR it is sufficient to know who usually makes decisions regarding large household purchases. Looking at the BN of CDR, *DecBigHousPurch* is linked to the BN by *Age*, nevertheless this arc is not robust, as shown in Table 1, suggesting that the other variables, with the exception of *GoOut*, could not be able to influence who usually makes decisions regarding large household purchases. At the contrary, looking at the BN of Egypt, all the variable can directly or indirectly influence who usually makes decisions regarding large household purchases. The MB of *DecBigHousPurch* is formed by *GoOut* and *Wealth*, so we can say that the knowledge of the attitude towards domestic violence and the wealth of the household in which the woman lives are sufficient to predict who usually makes decisions regarding large household purchases. Table 2 reports the MBs

for the variable *EducLev*. The two MB have nodes in common (*Wealth*, *Urban* and *AgeFrsMar*), and nodes that are specific for the two countries.

Table 2. *Markov blanket* for the variable *EducLev*.

Variables	Wealth	Urban	AgeFrsMar	IdealKid	Cheb	Work	Age	GoOut
CDR	✓	✓	✓	✓	✓			
Egypt	✓	✓	✓			✓	✓	✓

For CDR, the specific node takes into account the number of ideal kids and the number of children ever born, and it might be related to the vision of the woman who basically generates children, typical of a poor and less developed country. For Egypt the MB includes, as specific nodes, the age of the woman, the attitude towards domestic violence and the working status, aspects which might be considered important in the female empowerment of a more developed country and ahead in the demographic transition.

In conclusion, the use of BNs allowed us to highlight peculiarities in the relationship between the considered variables, suggesting the usefulness of the BN also in the comparison task. Future work will comprise the analysis of more African countries in order to find common and peculiar features.

References

1. Ewerling, F., Lynch, J.W., Victora, C.G., van Eerdekijk, A., Tysler, M., Barros, A.J.D.: The SWPER index for women's empowerment in Africa: development and validation of an index based on survey data. *Lancet Glob. Health* **5**, e916–e923 (2017)
2. Kabeer, N.: Resources, agency, achievements: reflections on the measurement of women's empowerment. *Dev. Chang.* **30**(3), 435–464 (1999)
3. Kabeer, N.: The Conditions and Consequences of Choice: Reflections on the Measurement of Women's Empowerment. UNRISD, Geneva (1999)
4. Kjerulff, U.B., Madsen, A.L.: Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, 2nd edn. Springer, New York (2013)
5. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997)
6. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
7. Malhotra, A., Schuler, S.R., Boender, C.: Measuring women's empowerment as a variable in international development. World Bank Workshop on Poverty and Gender: New Perspectives 28 (2002)
8. Scutari, M.: Learning bayesian networks with the bnlearn R package. *J. Stat. Softw.* **35**(3), 1–22 (2010)
9. Scutari, M., Denis, J.-B.: Bayesian Networks with Examples in R. CRC Press, Taylor & Franics Group (2015)



Inverse Stereographic Gaussian DAG Models

Anna Gottard^(✉) and Agnese Panzera

Department of Statistics, Computer Science, Applications “G. Parenti”,
University of Florence, Firenze, Italy
{anna.gottard, agnese.panzera}@unifi.it

Abstract. Directed acyclic graphical models are statistical models that use a directed acyclic graph (DAG) to represent the conditional independence structure of a set of random variables. We introduce a new class of DAG models tailored for circular variables having an Inverse Stereographic Normal distribution. We present a case study where the proposed models are employed to describe conditional independences of a sequence of angles summarising the structure of a protein.

Keywords: Bayesian networks · Circular distributions · Proteins structure · Toroidal data

1 Introduction

Circular observations are observations coming in the form of directions or angles and their sample space is the unit circle $\mathbb{T}$. Classical examples are the directions of migration of animal species, the directions of wind and marine currents, but also the time of the day at which a given phenomenon is recorded. The special nature of circular observations lies in their periodicity, which sets apart *circular statistics* from standard statistical methods conceived for Euclidean data. In the last decades, circular statistics has greatly evolved and many statistical methods now boast their circular counterparts. For complete overviews on statistics for circular data, see [3, 8] and for some collections of more recent contributions see [6, 7, 9].

In some context, the interest lies in studying the relations among p angles, which can be viewed as points on the surface of a p -dimensional torus, $\mathbb{T}^p$, obtained by the p -fold Cartesian product of unit circles. A classical example of toroidal data arises in structural bioinformatics problems involving the study of the relations among dihedral angles occurring in amino acid sequences of proteins structures. Despite the advances in circular statistics, the study of distributions on $\mathbb{T}^p$ and models involving p circular variables appears to be limited in the literature when $p > 2$. Furthermore, for example, some multivariate distributions pose issues of intractability of the likelihood function (as the wrapped Normal distribution) or the lack of closed-form for the normalization constant (as the multivariate von Mises distribution), which make inference more cumbersome.

Conversely, the multivariate Inverse Stereographic Normal distribution (see, for example, [12]) shares many desirable properties with the multivariate Normal distribution.

Graphical models (see, for example, [4]) are a class of probabilistic models for analyzing the conditional independence structure of a set of random variables. This class of multivariate models represents the conditional independence structure of p variables through a graph $\mathcal{G} = (V, E)$, where $V = \{1, \dots, p\}$ denotes a set of nodes and $E \subset V \times V$ is a set of edges. The nodes within V correspond to the random variables, while the absence of edges in E indicates conditional independence relationships. The joint probability distribution associated with the model factorizes with respect to $\mathcal{G}$. A graphical model is termed a *directed acyclic graph* (DAG) model or a *Bayesian network* when $\mathcal{G}$ is a directed acyclic graph containing only directed edges, namely $E \subseteq \{(i, j) : i, j \in V, i \neq j\}$ represents a collection of ordered edges, with no cycles permitted. Bayesian networks presume a certain ordering of the variables under study and offer a useful factorisation in terms of univariate conditional distributions related to the chain rule of probability. Except for some rather recent contributions, DAG models for circular variables seem to be a nearly unexplored topic. The first proposal, up to our knowledge, is provided by [1]. They studied the sequence and the structure of a protein, by using two classes of models combining Hidden Markov Models, interpreted as Bayesian networks, with distributions defined on the sphere and on the bidimensional torus. More recently, [5] considered the bivariate wrapped Cauchy distribution, which can be associated with a tree-structured directed acyclic graph. The assumed distribution, being defined on $\mathbb{T}^2$, turns out to be specific for graphs with a tree structure. [2] presented a class of DAG models related to the univariate von Mises distribution.

In this work, we propose a class of directed acyclic graph models based on the Inverse Stereographic Normal distribution on the torus. Due to the properties of the Inverse Stereographic Normal distribution, these models benefit of all the existing inferential procedures developed in the literature for the Gaussian Bayesian network.

The paper is organized as follows. Section 2 introduce the Inverse Stereographic DAG model, while Sect. 3 is devoted to apply the proposed class of models to describe the conditional independence structure of the angles in the structure of a reovirus protein. Finally, Sect. 4 contains some concluding remarks.

2 A DAG Model for Circular Data

The probability density function of a *circular* random variable Θ , say f_Θ , is defined on $\mathbb{T} := [0, 2\pi)$ and satisfies $f_\Theta(\theta) = f_\Theta(\theta + 2\pi z)$, for any $\theta \in \mathbb{T}$ and any $z \in \mathbb{Z}$. There are various approaches to define circular distributions, one of which is based on the idea of considering a *linear* distribution for the stereographic projection of Θ . The family of Inverse Stereographic Normal (ISN) distributions is obtained using this approach. Letting $U = \tan(\Theta/2)$ and assuming $U \sim N(\mu, \sigma^2)$, Θ has an Inverse Stereographic Normal distribution, namely $\Theta \sim$

$ISN(\mu, \sigma^2)$. For the general p -variate case, letting $\text{Sym}_+^{p \times p}$ denote the set of symmetric positive definite matrices of order p , we say that a vector of random angles $\boldsymbol{\Theta} = (\Theta_1 \dots \Theta_p)'$ has an Inverse Stereographic Normal distribution, $\boldsymbol{\Theta} \sim ISN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \text{Sym}_+^{p \times p}$, if its probability density function at a point $\boldsymbol{\theta} = (\theta_1 \dots \theta_p)' \in \mathbb{T}^p$ is

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} \exp((\mathbf{u} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu})) \prod_{j=1}^p \frac{1}{1 + \cos(\theta_j)},$$

where $\mathbf{u} = (u_1 \dots u_p)'$ with $u_j = \tan(\theta_j/2)$ being the *stereographic projection* of $(\cos(\theta_j) \ \sin(\theta_j))'$. If $\mathbf{U} = \tan(\boldsymbol{\Theta}/2)$ and $\boldsymbol{\Theta}$ follows an $ISN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, then $\mathbf{U}$ follows a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then, the distribution of $\boldsymbol{\Theta}$ shares the location and scale parameters with the distribution of $\mathbf{U}$. Moreover, due to its formulation, the ISN distribution shares with the Gaussian distribution the properties of closure under both marginalization and conditioning.

For a DAG $\mathcal{G}$ with vertex set $V = \{1, \dots, p\}$, let $\text{pa}(j)$ be the set of parents of the node $j \in V$, that is the set of vertices such that there is an arrow from them to j . Assume that the elements of V are *well-ordered* in the sense that they are linearly ordered in a way which is compatible with $\mathcal{G}$, i.e. so that $i \in \text{pa}(j)$ implies $i < j$. Moreover, for a p -dimensional vector $\boldsymbol{\Theta}$ and for a subset A of V , let the vector containing the variables indexed in A . Then we get the following definition.

Definition 1. (Inverse Stereographic Normal directed acyclic graphical model) Let $\mathcal{G} = (V, E)$ be a directed acyclic graph with ordered set of vertices $V = (1, \dots, p)$ and set of directed edges $E \subset V \times V$. The Inverse Stereographic Gaussian DAG model over $\mathcal{G}$, denoted by $ISN_{\mathcal{G}}$, consists of all multivariate Inverse Stereographic Normal distributions which obey the directed Markov property with respect to $\mathcal{G}$, i.e. if

$$\boldsymbol{\Theta} = (\Theta_1 \dots \Theta_p)' \sim ISN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in ISN_{\mathcal{G}},$$

then, for each j ,

$$\Theta_j \perp\!\!\!\perp \boldsymbol{\Theta}_{\{1, \dots, j-1\} \setminus \text{pa}(j)} \mid \boldsymbol{\Theta}_{\text{pa}(j)}.$$

The above property is equivalent to the following factorisation of the joint distribution of $\boldsymbol{\Theta}$. For generic subsets A, B of V , respectively denote as $\boldsymbol{\mu}_A$ and $\boldsymbol{\Sigma}_A$ the mean vector and the covariance matrix of the random vector $\mathbf{U}_A$, and denote as $\boldsymbol{\Sigma}_{AB}$ the part of the covariance matrix $\boldsymbol{\Sigma}$ containing the covariances between $\mathbf{U}_A$ and $\mathbf{U}_B$. Then if $\boldsymbol{\Theta} = (\Theta_1, \dots, \Theta_p) \sim ISN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $ISN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in ISN_{\mathcal{G}}$,

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \prod_{j=1}^p f_{\Theta_j \mid \boldsymbol{\Theta}_{\text{pa}(j)}}(\theta_j \mid \boldsymbol{\theta}_{\text{pa}(j)}),$$

where

$$\Theta_j \mid \boldsymbol{\Theta}_{\text{pa}(j)} \sim ISN(\mu_{j \mid \text{pa}(j)}, \Sigma_{j \mid \text{pa}(j)}),$$

with

$$\mu_{j|pa(j)} = \boldsymbol{\mu}_j + \boldsymbol{\Sigma}_{jpa(j)} \boldsymbol{\Sigma}_{pa(j)}^{-1} (\boldsymbol{\theta}_{pa(j)} - \boldsymbol{\mu}_{pa(j)}),$$

and, denoting $\boldsymbol{\Sigma}_{pa(j)}^T$ the transpose of $\boldsymbol{\Sigma}_{jpa(j)}$,

$$\boldsymbol{\Sigma}_{j|pa(j)} = \boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}_{jpa(j)} \boldsymbol{\Sigma}_{pa(j)}^{-1} \boldsymbol{\Sigma}_{pa(j)}^T.$$

For Inverse Stereographic Normal DAG models, both the estimation and the edge selection can be carried out by using classical approaches for Gaussian DAG models. Indeed, the factor which diversifies the densities of the ISN and Normal distributions does not involve any parameter. A drawback of the ISN distribution is the lack of invariance under rotation. Thus, the definition of the model is conditional on the choice of the origin. For further details, see [12].

3 The Reovirus Fusion-Associated Small Transmembrane Protein

We present an illustrative example where the reovirus fusion-associated small transmembrane protein, whose structure has been measured via NMR solution by [10]. The data set is available at the RCSB Protein Data Bank under the name of 2MNS. Reoviruses are a family of viruses that can affect the human gastrointestinal system and respiratory tract. The 2MNS protein (see Fig. 1) has a chain structure with 22 residue counts, with a total number of 40 dihedral angles, measured on $n = 50$ models from Baboon orthoreovirus by [10].

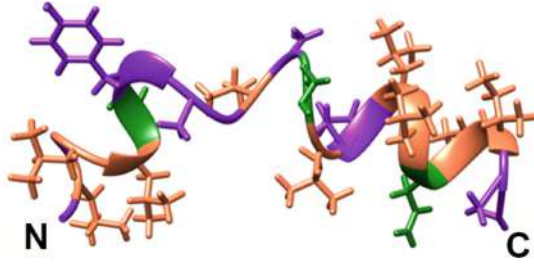


Fig. 1. NMR 3D structure determination of 2MNS protein by [10]

Assuming as natural ordering of the dihedral angles of the 2MNS protein its primary structure, we use the Incremental Association algorithm to select the DAG edges as proposed in `bnlearn` [11]. This algorithm is based on the Markov blanket detection algorithm and uses forward selection with false positives removal [13]. Figure 2 depicts the resulting graph, highlighting the different peptides by colors and numbering the pair of dihedral angles ϕ and ψ according to their place in the primary structure sequence. It can be noted that the association follows the 3D structure (Fig. 1) with some further connections apparently linked to the presence of the helix.

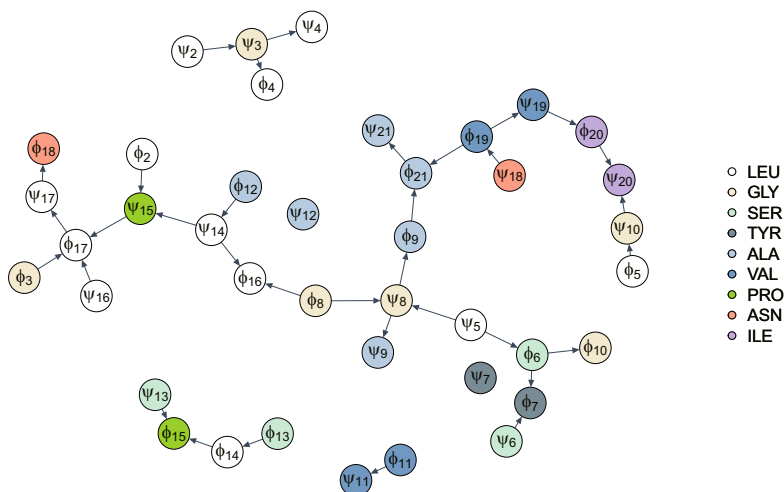


Fig. 2. Selected DAG for Reovirus dihedral angles ϕ and ψ , assuming the primary structure as natural nodes ordering. Colors distinguish peptides and numbers denote the ordering.

4 Concluding Remarks

We introduce a novel class of Directed Acyclic Graph models tailored for analyzing the independence structures among circular random variables. This class of models draws on the multivariate Inverse Stereographic Normal distribution. This is a toroidal distribution defined according the *embedding* approach for formulating probability distributions of random angles. Remarkably, this distribution maintains the desirable properties of closure under marginalization and conditioning, akin to those of the multivariate Normal distribution. The proposed class of models takes advantage of the well-established procedures developed for the class of Gaussian Bayesian network. Furthermore, this class of models shows great promise for various applications, notably in structural bioinformatics, among others.

References

1. Boomsma, W., Kent, J.T., Mardia, K.V., Taylor, C.C., Hamelryck, T.: Graphical models and directional statistics capture protein structure. *Interdisc. Stat. Bioinform.* **25**, 91–94 (2006)
2. Gottard, A., Panzera, A.: Graphical models for circular variables. *arXiv preprint [arXiv:2104.03194](https://arxiv.org/abs/2104.03194)* (2021)
3. Jammalamadaka, S.R., Sengupta, A.: *Topics in Circular Statistics*, vol. 5. World Scientific, Singapore (2001)
4. Lauritzen, S. L.: *Graphical Models*, vol. 17. Clarendon Press, Oxford (1996)

5. Leguey, I., Bielza, C., Larrañaga, P.: Tree-structured Bayesian networks for wrapped Cauchy directional distributions. In: Luaces, O., et al. (eds.) CAEPIA 2016. LNCS (LNAI), vol. 9868, pp. 207–216. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44636-3_19
6. Ley, C., Verdebout, T.: Modern Directional Statistics. CRC Press, Boca Raton (2017)
7. Ley, C., Verdebout, T.: Applied Directional Statistics: Modern Methods and Case Studies. CRC Press, Boca Raton (2018)
8. Mardia, K.V., Jupp, P.E.: Directional Statistics, vol. 494. Wiley, Chichester (2009)
9. Pewsey, A., García-Portugués, E.: Recent advances in directional statistics. TEST **30**(1), 1–58 (2021). <https://doi.org/10.1007/s11749-021-00759-x>
10. Read, J., et al.: Reovirus FAST proteins drive pore formation and syncytiogenesis using a novel helix-loop-helix fusion-inducing lipid packing sensor. PLoS Pathogens **11**(6), e1004962 (2015)
11. Scutari, M.: Learning Bayesian networks with the BNlearn R package. J. Stat. Softw. **35**(3), 1–22 (2010)
12. Selvitella, A.: On geometric probability distributions on the torus with applications to molecular biology. Electron. J. Stat. **13**(2), 2717–2763 (2019)
13. Tsamardinos I., Aliferis C.F., Statnikov A.: Algorithms for large scale Markov blanket discovery. In: Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference, pp. 376–381 (2003)



Addressing Self-selection in Long Covid Studies: A Methodological Framework

Konstantina Gourgoura¹, Elena Stanghellini¹, and Marco Doretti²(✉)

¹ Department of Economics, University of Perugia, via A. Pascoli 20, 06122 Perugia, Italy

² Department of Statistics, Computer Science, and Applications, University of Florence, viale Morgagni 59, 50134 Florence, Italy
`marco.doretti@unifi.it`

Abstract. A methodological framework to address self-selection is presented that combines previous results on stratified case-control sampling designs with a newly developed theory on secondary outcome analyses. The derivations are then applied to data coming from a single-centre longitudinal prospective cohort study on Post and Long Covid syndrome, conducted at Luigi Sacco University Hospital in Milan from May 2020 to October 2022.

Keywords: distortion · M -estimation · logistic model · sampling bias · selection bias · self-selection

1 Introduction

Self-selection into studies is a common problem in epidemiology, giving rise to a sampling bias that cannot be resolved by simply increasing the number of observations. Recent review on the subject can be found in [7].

Typical epidemiological contexts where these issues emerge are studies on Post/Long Covid (PLC) syndrome. PLC syndrome is a direct sequelae of SARS-CoV-2 virus that can highly compromise the quality of life. It is composed of (1) symptoms that persist from the acute COVID-19 phase or its treatment, (2) symptoms that have resulted in a new health limitation, (3) new symptoms that have occurred after the end of the acute phase but are understood to be a consequence of COVID-19 disease, (4) worsening of a pre-existing underlying condition; see [2].

In order to understand the dynamics of PLC, several studies have been designed. Typically, participants are recruited by simply advertising the study, and people enter on a voluntary base. This creates a possible bias, as people who are experiencing lasting symptoms after COVID-19 may be more motivated than others to participate. Moreover, patients who are not recovering from the syndrome are more likely to participate at follow-up, giving rise to informative drop-out; see [6], where it is claimed that the analysis of the determinants is biased and the real prevalence of people affected by PLC syndrome is unknown.

After presenting a methodological framework to address the issue of self-selection, we analyse data coming from a single-centre longitudinal prospective cohort study conducted at Luigi Sacco University Hospital (LSUH) in Milan from May 2020 to October 2022. Hospitalised and non-hospitalised patients older than 18, after signing a written informed consent, were enrolled in a clinical and immunological longitudinal study approved by the competent ethical committee (“Comitato Etico Internazionale”, Area 1, n. 2020/ST/158). Patients were either referred by specialists, general practitioners or self-referred. We focus on the patients from the first wave (i.e. with date of COVID-19 onset of symptoms before 01/10/2020). Several symptoms were investigated (palpitations, amnesia, headache, anxiety/panic, insomnia, loss of smell, loss of taste, dyspnoea, asthenia, myalgia, telogen effluvium and other different ones). A number of possible covariates (such as age, gender, comorbidities, obesity) are measured on the patients prior to infection. A measure of the severity of the COVID-19 is also available, according to WHO score into 4 categories: mild, moderate, severe and critical. A number of research questions are of interest, ranging from the effect of patients pre-infection conditions (such as underlying conditions and obesity) on the probability to develop COVID-19 into either severe or critical form to the effect of some treatments as well as pre-infection conditions on the probability to develop PLC syndrome. With reference to the role of obesity, we here present a modelling strategy that allows to overcome the issue of self-selection.

2 Addressing Self-selection: A Methodological Approach

The object of the investigation is the effect of a variable A (in our case, the obesity indicator) on either a response variable Y (in our case, the presence of at least one PLC symptom) or a response variable Z (in our case, whether or not the COVID-19 infection is severe or critical). For reasons that will become clearer in the next paragraph, we assume to have also two distinct sets of background covariates representing personal characteristics, denoted by B and C . In particular, the former includes categorical variables, whereas variables in the latter can be of any nature. In this setting, we let

$$\text{logit}\{P(Y = 1 \mid a, b, c, z)\} = \mathbf{x}_y^\top \boldsymbol{\beta} \quad (1)$$

and

$$\text{logit}\{P(Z = 1 \mid a, b, c)\} = \mathbf{x}_z^\top \boldsymbol{\delta}, \quad (2)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are coefficient vectors, while $\mathbf{x}_y$ and $\mathbf{x}_z$ contain the values of the explanatory variables (including the intercept and A). Without loss of generality, we assume that all variables in (B, C) influence both the outcome Y and the outcome Z . Clearly, such an assumption can be removed in that some of them may not be relevant predictors of Y or Z , respectively.

To address the issue of self-selection of patients into the analysis, we introduce a selection indicator variable, W , such that data are available just on the selected units, i.e. those with $W = 1$. We assume that self-selection is mainly driven by

Y . This means that W is probabilistically influenced by Y either marginally or conditionally on the strata formed by the set of categorical background covariates, B . We remark that this is the mechanism operating also in (stratified) case-control studies designed for Y , typically the outcome of primary interest. As data are available on patients with $W = 1$ only, a distortion is induced in both marginal models, as:

$$\begin{aligned}\text{logit}\{P(Y = 1 \mid a, b, c, z, W = 1)\} &\neq \mathbf{x}_y^\top \boldsymbol{\beta} \\ \text{logit}\{P(Z = 1 \mid a, b, c, W = 1)\} &\neq \mathbf{x}_z^\top \boldsymbol{\delta}.\end{aligned}$$

3 The Impact of Self-selection on the Equation for Y

It is well-known that simple coefficient adjustments involving the conditional population prevalences $\pi_b = P(Y = 1 \mid B = b)$ suffice. In detail, following notation in [3] we have that the model expression holding for the selected units is

$$\text{logit}\{P(Y = 1 \mid a, b, c, z, W = 1)\} = \mathbf{x}_y^\top \boldsymbol{\beta}^*, \quad (3)$$

where $\boldsymbol{\beta}^*$ is equal to $\boldsymbol{\beta}$ except for modifications concerning the intercept (β_{INT}) and the coefficient sub-vector $\boldsymbol{\beta}_B = (\beta_2, \dots, \beta_{N_B})^\top$ related to the $N_B - 1$ stratum indicator variables. Specifically, letting $p_b = P(Y = 1 \mid W = 1, B = b)$ be the sample prevalence within the b -th stratum, the replacing elements in $\boldsymbol{\beta}^*$ are

$$\begin{aligned}\beta_{\text{INT}}^* &= \beta_{\text{INT}} + \log(k_1) \\ \boldsymbol{\beta}_B^* &= \boldsymbol{\beta}_B + \log(\mathbf{k}_{-1}),\end{aligned} \quad (4)$$

where $\mathbf{k}_{-1} = k_1^{-1} \cdot (k_2, \dots, k_{N_B})^\top$ and, for $b \in \{1, \dots, N_B\}$,

$$\begin{aligned}k_b &= \frac{P(W = 1 \mid Y = 1, B = b)}{P(W = 1 \mid Y = 0, B = b)} \\ &= \frac{P(Y = 1 \mid W = 1, B = b)}{P(Y = 0 \mid W = 1, B = b)} \cdot \frac{P(Y = 0 \mid B = b)}{P(Y = 1 \mid B = b)} = \frac{p_b}{1 - p_b} \cdot \frac{1 - \pi_b}{\pi_b}.\end{aligned}$$

4 The Impact of Self-selection on the Equation for Z

We now turn to the model for $P(Z = 1 \mid a, b, c, y, W = 1)$. If, after conditioning on (A, B, C, Y) , Z and W can be assumed to be independent, then $P(Z = 1 \mid a, b, c, y, W = 1) = P(Z = 1 \mid a, b, c, y)$. Then, after some derivations, see Doretti et al. (2024) Appendix A, it is possible to see that

$$\text{logit}\{P(Z = 1 \mid a, b, c, y)\} = \mathbf{x}_z^\top \boldsymbol{\delta} + o(y, \mathbf{x}_y; \boldsymbol{\beta}), \quad (5)$$

where

$$o(y, \mathbf{x}_y; \boldsymbol{\beta}) = \log \frac{P(Y = y \mid a, b, c, Z = 1)}{P(Y = y \mid a, b, c, Z = 0)} \quad (6)$$

is a correction term depending on y and the coefficient vector $\boldsymbol{\beta}$. The parametric expression of it is in [3]. It then follows that in this second case, Y acts as a *selection bias breaking node*. If Y is binary and π_b is known, the logistic model for Z can be identified. M -estimation can be performed via simply offsetting the $o(y, \mathbf{x}_y; \boldsymbol{\beta})$ term; see [3] where also ML is implemented.

5 The Role of Obesity in the LSUH Study

As mentioned earlier, we apply the above theoretical framework to the first data wave of the LSUH cohort study; see [1] for details about this study. Y is the binary variable indicating whether at least one of the PLC symptoms appeared within 2 months from the end of COVID-19 and lasted for more than 2 months (see WHO definition). For the data at hand, we assume that self-selection is induced by Y only. After some preliminary data cleansing, 362 patients were included in the final dataset, 278 of which with $Y = 1$. The effect of obesity (encoded by the binary indicator A) is investigated with respect to Y as well as to Z , a marker indicating whether the COVID-19 infection is severe or critical (see Sect. 2).

Table 1. Estimated logistic regression model for Y , prevalence $\pi = 0.10$.

	estimate	standard error	p-value
Intercept	-3.442	0.472	0.000
$\mathbb{I}\{A = 1\}$	0.568	0.469	0.226
$\mathbb{I}\{\text{Gender} = \text{Female}\}$	0.688	0.266	0.010
Age	0.012	0.009	0.191
$\mathbb{I}\{Z = 1\}$	0.908	0.321	0.005

In Table 1, the output of the logistic model for Y given A , Z as well as age and gender is reported. Since by assumption no variables other than Y drive the self-selection mechanism, the B set is empty and the only needed adjustment concerns the intercept, with no impact on the interpretation of the explanatory variable effects. In detail, results in the table refer to the population prevalence level $\pi = 0.10$. The model includes main effects only, since no significant interaction was found.

While the effect of obesity is positive but not significant, the effect of Z is both positive and significant, showing that, as expected, patients with more severe COVID-19 infections are more likely to develop at least one PLC symptom. As for the adjusting covariates, the effect of age is positive but not significant, whereas the effect of gender is significant, with females being more likely to develop - other covariates being constant - the PLC syndrome.

With regard to the model for Z , we recall that the prevalence of PLC syndrome in the population of those who recovered from the infection is necessary in order to use the derivations in previous section (that is, in order to build the adjusting offset term in (6)). However, as noticed in [6], due to self-selection issues the real value of this prevalence is unknown. A sensitivity analysis approach is therefore of interest. With this respect, in Fig. 1 we plot the estimated coefficient of the obesity indicator, A , as a function of the PLC prevalence, with the latter varying from 0.05 to 0.50. The considered model includes also age,

gender and a binary indicator of whether patients present a renal disease at baseline. While the relationship exhibits a quadratic behaviour, it is worth to mention that for every prevalence level the estimated effect of A is significant, since standard errors are always very close to 0.35, while the estimated coefficients lie between 1.18 and 1.27. In the graph, the dashed line represents the effect (with value 1.21) estimated by the naive model, not including the offset correction term.

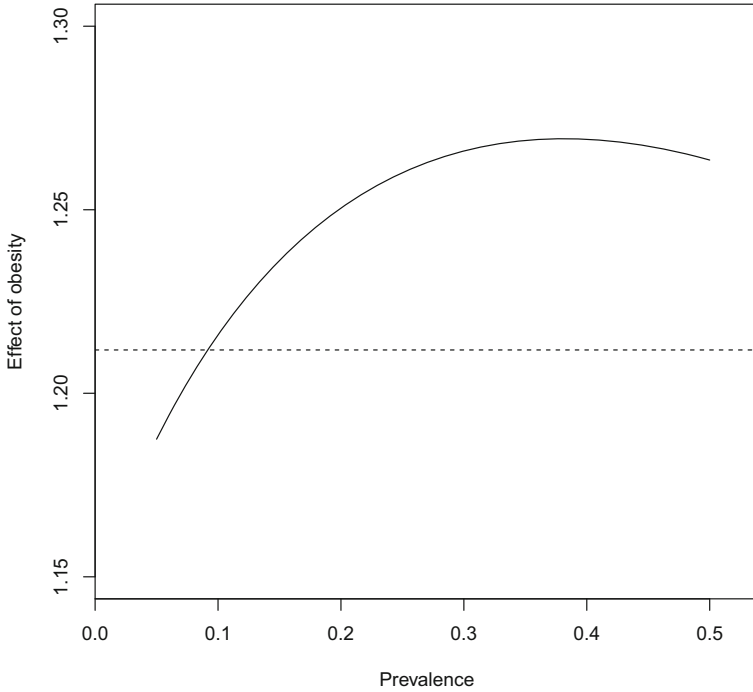


Fig. 1. Estimated effect of obesity on Z as a function of the prevalence of the PLC syndrome. The dashed line corresponds to the effect estimated with the unadjusted model.

6 Conclusions

We presented a methodological framework to address the issue of self-selection of patients into studies. Focus on other aspects of the syndrome is ongoing, also with respect to treatment strategies and different waves. Notice that other parametric strategies exist that model explicitly the self-selection mechanism, see [4], where the extension of Heckman's model to two time occasions has been presented. Notice further that when point identification is not possible, results on interval identification have been proposed, see e.g. [5].

Acknowledgements. Support by “Fondazione Perugia”, project “Danni permanenti dell’infezione da Sars-Cov-2”, is gratefully acknowledged. We thank Dr. Fabio Borgonovo and Prof. Paolo Berta for assistance on medical issues. Dr. Perla Rivadeneyra is also gratefully acknowledged for managing and adjusting the original database.

References

1. Borgonovo, F., et al.: Analysis and clinical determinants of post-COVID-19 syndrome in the Lombardy region: evidence from a longitudinal cohort study. *BMJ Open* **14**(2), e075185 (2024)
2. Cabrera Martimbianco, A.L., Pacheco, R.L., Bagattini, Â.M., Riera, R.: Frequency, signs and symptoms, and criteria adopted for long COVID-19: a systematic review. *Int. J. Clin. Pract.* **75**(10), e14357 (2021)
3. Doretto, M., Genbäck, M., Stanghellini, E.: Mediation analysis with case-control sampling: identification and estimation in the presence of a binary mediator. *Biom. J.* **66**(1), 2300089 (2024)
4. Doretto, M., Geneletti, S., Stanghellini, E.: Tackling non-ignorable dropout in the presence of time varying confounding. *J. R. Stat. Soc.: Ser. C: Appl. Stat.* **65**(5), 775–795 (2016)
5. Genbäck, M., Stanghellini, E., de Luna, X.: Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome. *Stat. Pap.* **56**, 829–847 (2015)
6. Høeg, T.B., Ladhani, S., Prasad, V.: How methodological pitfalls have created widespread misunderstanding about long COVID. *BMJ Evidence-Based Med.* (2023). <https://doi.org/10.1136/bmjebm-2023-112338>
7. Mathur, M.B., Shpitser, I.: Reflections on evolving conceptions of selection bias. Technical report, Center for Open Science (2024)



An R-Shiny Application for Statistical Matching

Fedele Greco, Lucia Guastadisegni^(✉), and Carlo Trivisano

Department of Statistical Sciences, University of Bologna, Bologna, Italy
{fedele.greco,lucia.guastadisegni2,carlo.trivisano}@unibo.it

Abstract. Statistical matching involves integrating different data sources with different units and a set of common variables. Applying statistical matching methods presents challenges for practitioners, including data pre-processing and the application of different statistical matching methods. To address this, we develop an R-Shiny application to simplify the integration of different datasets for users and enhance efficiency in statistical matching tasks. In this paper, we delineate the various features of the application, encompassing tabs for importing new data or selecting from available surveys, choosing the matching variable, conducting data wrangling, employing diverse statistical matching methods, and using the matched dataset for regression analysis.

Keywords: Complex data survey · data integration · interactive tools · statistical matching

1 Introduction

Statistical matching encompasses a class of statistical methods designed to combine different datasets that share a set of common variables, some variables that are not jointly observed, across disjoint sets of units [2]. A notable characteristic of statistical matching methods is that they allow us to explore relationships between variables that are not observed simultaneously. Integration of two or more data sources can follow either a micro approach, where the objective is the construction of a complete synthetic file, or a macro approach, where the objective is the estimation of the joint distribution function of the variables of interest that have not been observed jointly [2]. Practitioners dealing with the integration of diverse data sources face numerous challenges, including recoding or recategorization of shared variables, substitution of some variables with new ones, conversion of measurement units, alignment of reference periods, and handling of missing or inconsistent data [2]. Moreover, applying various statistical matching techniques and generating a synthetic dataset can be intricate tasks. Therefore, the objective of this work is to develop an R-Shiny [4] application to aid practitioners in all stages of statistical matching, from importing raw data to creating matched synthetic datasets. The application will also facilitate regression analysis using the matched data. Additionally, it will provide R code for

generating matched datasets from raw data, ensuring the reproducibility of the analysis.

The paper is organized as follows. In Sect. 2, we offer a concise review of prominent statistical matching methods. Section 3 outlines the software description, highlighting its diverse features. Finally, in Sect. 4, we present and discuss some concluding remarks.

2 Statistical Framework and Methods

Let Y and Z be two random variables (r.v.s) and $\mathbf{X} = (X_1, \dots, X_P)^\top$ be a vector of r.v.s of dimension P . Consider two datasets A and B referred to n_A and n_B statistical units, respectively. For each unit, independent and identically distributed (i.i.d.) observations generated from the joint distribution $f(\mathbf{x}, y, z)$ are available. The units in A have Z missing and the units in B have Y missing, that is:

$$(\mathbf{x}_a^A, y_a^A)^\top = (x_{a1}^A, \dots, x_{aP}^A, y_a^A)^\top, \quad (\mathbf{x}_b^B, z_b^B)^\top = (x_{b1}^B, \dots, x_{bP}^B, z_b^B)^\top, \quad (1)$$

where $a = 1, \dots, n_A$ and $b = 1, \dots, n_B$. Statistical matching operates under the hypothesis that A and B are sampled from the same target population. We consider statistical matching methods based on the *Conditional Independence Assumption* (CIA) between Y and Z given $\mathbf{X}$, with density [2]

$$f(\mathbf{x}, y, z) = f_{Y|\mathbf{X}}(y|\mathbf{x}) \cdot f_{Z|\mathbf{X}}(z|\mathbf{x}) \cdot f_{\mathbf{X}}(\mathbf{x}). \quad (2)$$

The CIA cannot be tested but makes the model identifiable. Some of the methods that can be adopted under the CIA are described in the next subsections. As a general rule for all methods, A is the recipient file and B is the donor file. Usually, $n_B \geq n_A$. The final matched dataset has n_A units and includes the $\mathbf{x}$, y , and $\tilde{z}$ variable values, where $\tilde{z}$ has been imputed.

2.1 Parametric Methods

Parametric methods assume a parametric form for (2), that is

$$f(\mathbf{x}, y, z; \boldsymbol{\theta}) = f_{Y|\mathbf{X}}(y|\mathbf{x}; \boldsymbol{\theta}_{y|\mathbf{x}}) \cdot f_{Z|\mathbf{X}}(z|\mathbf{x}; \boldsymbol{\theta}_{z|\mathbf{x}}) \cdot f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}_{\mathbf{x}}), \quad (3)$$

where $\boldsymbol{\theta}_{y|\mathbf{x}} \in \Theta_{y|\mathbf{x}}$, $\boldsymbol{\theta}_{z|\mathbf{x}} \in \Theta_{z|\mathbf{x}}$, $\boldsymbol{\theta}_{\mathbf{x}} \in \Theta_{\mathbf{x}}$. Concerning the macro approach, thanks to the CIA, the model is identifiable, and maximum likelihood (ML) estimates of the parameters can be obtained by maximizing the following log-likelihood [9]:

$$\begin{aligned} L(\boldsymbol{\theta} | \mathbf{A} \cup \mathbf{B}) &= \prod_{a=1}^{n_A} f_{\mathbf{X}Y}(\mathbf{x}_a, y_a; \boldsymbol{\theta}) \prod_{b=1}^{n_B} f_{\mathbf{X}Z}(\mathbf{x}_b, z_b; \boldsymbol{\theta}) \\ &= \prod_{a=1}^{n_A} f_{Y|\mathbf{X}}(y_a | \mathbf{x}_a; \boldsymbol{\theta}_{Y|\mathbf{X}}) \prod_{b=1}^{n_B} f_{Z|\mathbf{X}}(z_b | \mathbf{x}_b; \boldsymbol{\theta}_{Z|\mathbf{X}}) \\ &\quad \times \prod_{a=1}^{n_A} f_{\mathbf{X}}(\mathbf{x}_a; \boldsymbol{\theta}_{\mathbf{X}}) \prod_{b=1}^{n_B} f_{\mathbf{X}}(\mathbf{x}_b; \boldsymbol{\theta}_{\mathbf{X}}). \end{aligned} \quad (4)$$

Conditional Mean Matching. After estimating (4), a complete synthetic dataset can be reconstructed. If Z is continuous

$$\tilde{z}_a = E(Z|X = \mathbf{x}_a) = \int_Z z f_{z|x}(z|\mathbf{x}_a; \boldsymbol{\theta}_{z|x}) dz \quad a = 1, \dots, n_a. \quad (5)$$

When Z is normally distributed, this is a classic regression imputation method [6], that is

$$\tilde{z}_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} \mathbf{x}_a^A \quad a = 1, \dots, n_a. \quad (6)$$

Draws Based on Conditional Predictive Distributions. A limitation of conditional mean matching is that it lacks variability when imputing values for observations with identical conditioning variables. To address this limitation, random values can be drawn from $f_{Z|X}(z|\mathbf{x}_a; \hat{\boldsymbol{\theta}}_{z|X})$ for every $a = 1, \dots, n_A$. When Z is normal this process is known as stochastic regression imputation [6]:

$$\tilde{z}_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} \mathbf{x}_a^A + \epsilon_a \quad a = 1, \dots, n_a, \quad (7)$$

where ϵ_a is randomly generated from a normal with 0 mean and estimated residual variance $\hat{\sigma}_{Z|X}^2$.

2.2 Non-parametric Methods

To avoid misleading results based on incorrect parametric assumptions, some non-parametric methods have been proposed for statistical matching, both at the micro and macro level. We consider only methods at the micro level, also known as hot-deck methods [12].

Random Hot Deck Method. This method randomly selects a donor record from the donor file for each record in the recipient file. Sometimes, this selection is done within specific subsets of units within the donor files using one or a few categorical variables shared between both files, such as demographic variables, forming what are known as donation classes [12].

Rank Hot Deck Method. In the rank hot deck method, a single ordinal matching variable X can be used for selecting donors to assign to the records in A [11]. Units in both files are ranked separately based on the values of X . If $n_B = kn_A$, where k is an integer, files are matched by pairing records with the same rank. In cases where the sample sizes of A and B differ, matching is conducted through the empirical cumulative distribution function of the distribution of X in the recipient file and donor file

$$\hat{F}_A^X(x) = \frac{1}{n_A} \sum_{a=1}^{n_A} I(x_a \leq x), \quad x \in X, \quad \hat{F}_B^X(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(x_b \leq x), \quad x \in X. \quad (8)$$

Subsequently, each $a = 1, \dots, n_A$ is linked to the corresponding record b^* in B such that ([2])

$$\left| \hat{F}_A^X(x_a^A) - \hat{F}_B^X(x_{b^*}^B) \right| = \min_{1 \leq b \leq n_B} \left| \hat{F}_A^X(x_a^A) - \hat{F}_B^X(x_b^B) \right|.$$

Distance Hot Deck Method. In the unconstrained distance method, when there is a single continuous matching variable X , the donor for the a -th record in the recipient file A is chosen such that [8]

$$d_{ab^*} = |x_a^A - x_{b^*}^B| = \min_{1 \leq b \leq n_B} |x_a^A - x_b^B|. \quad (9)$$

In the constrained distance methods, each record in B can be chosen as donor only once ($n_a \leq n_b$).

The hot deck distance-based techniques offer versatility by enabling integration with a variety of metrics. For continuous variables, this includes the Manhattan, Mahalanobis, and Exact distances, while for variables of different natures, the Gower index is utilized.

2.3 Mixed Methods

Mixed methods employ a combination of parametric and non-parametric techniques. The initial step involves estimating a parametric model for $(\mathbf{X}, Y, Z)$, followed by the derivation of a complete 'synthetic' dataset using a non-parametric approach. If Y and Z are continuous, the method comprises the following steps [10]:

1. The regression parameters of Z on X are estimated on file B, and the regression parameters of Y on X are estimated on file A.
2. Intermediate values $\tilde{z}_a$ are computed as in Equation (6) or (7) for the units in A. Similarly, values $\tilde{y}_b$ are also computed for the units in B.
3. For each $a = 1, \dots, n_A$, impute the value z_b^* corresponding to the nearest neighbour b^* in B with respect to a distance between intermediate values determined at step 2, $d((y_a, \tilde{z}_a), (\tilde{y}_b, z_b))$.

Various versions of the aforementioned method have been proposed in the literature (see for example [10, 12]), each implying modifications to either point 2 or 3.

3 Software Description

The different features of the application can be accessed through a navigation sidebar that contains multiple tabs, each corresponding to a different phase of the statistical matching procedure. More in details, it is composed by the following tabs:

1. Summary of surveys and selection of the recipient and donor files: the first tab allows the user to select the surveys to be used for statistical matching. Some surveys are included in the application to enable the user to train with various statistical matching methods and explore the application's potential. Brief descriptions of each survey and variables are provided. Moreover, users have the opportunity to import new surveys into the application. The recipient file (A) and donor file (B) are selected.
2. Selection of the matching variables: once the A and B files have been selected in the first tab, the second tab allows the user to select the matching variables, i.e., the variables that are common to the two surveys. These variables, denoted as $\mathcal{I}_A$ and $\mathcal{I}_B$, are not yet ready to be used for statistical matching because they could exhibit inconsistencies between the two datasets.
3. Data wrangling and harmonization: in this tab, inconsistencies among the variables $\mathcal{I}_A$ and $\mathcal{I}_B$ can be resolved. In detail, categorical variables can be harmonized to have the same categories, continuous variables can be transformed into categorical ones (for example, age in years can be transformed into age classes), some variables can be transformed into new variables (for example, individual-level variables can be summarized into household-level variables), measurement errors can be adjusted, and so on. The harmonized variables are now denoted as $\mathcal{X}_A$ and $\mathcal{X}_B$.
4. Application of matching methods: in this tab, the application of statistical matching methods is enabled. Some of the statistical matching methods outlined in Sect. 2 are implemented using the StatMatch R package [3]. Within this tab, the variable Z is selected from dataset B, and the variable Y from dataset A. Through the different statistical matching methods presented in Sect. 2, the variable Z is imputed into dataset A. Several columns $\tilde{z}_a$ can be generated, each corresponding to a different matching method. This facilitates comparison among the various matching techniques.
5. Employment of matched data for estimating regression parameters: the last tab of the application allows the user to estimate regression models using the matched dataset. This enables the comparison of different regression parameters obtained from various matching methods [5].

The application allows downloading the R code needed to perform all the statistical analyses necessary to obtain matched data starting from raw data, thereby fostering reproducibility.

The application is currently under development, but it will be available by the time of the SIS Conference 2024.

4 Conclusion

In this work, we presented an R-Shiny application for statistical matching. The application assists the user in all phases of statistical matching procedures, from data importation to the application of regression analysis on matched data. Currently, the application is limited to statistical matching methods based on

the CIA. As future work, more recent methods based on auxiliary information [7] could be included in the application.

Acknowledgements. This study was funded by the European Union - NextGenerationEU, in the framework of the “GRINS -Growing Resilient, INclusive and Sustainable project” (PNRR - M4C2 - I1.3 - PE00000018 - CUP J33C22002910001). 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.

References

1. Burkard, R.E., Derigs, U.: Assignment and Matching Problems: Solution Methods with FORTRAN-Programs, vol. 184. Springer, Heidelberg (2013). <https://doi.org/10.1007/978-3-642-51576-7>
2. D’Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching: Theory and Practice. Wiley, New York (2006)
3. D’Orazio, M.: StatMatch: Statistical Matching or Data Fusion. R package version 1.4.1. (2022)
4. Chang, W, et al.: shiny: Web Application Framework for R, R package version 1.8.0.9000 (2024). <https://github.com/rstudio/shiny>, <https://shiny.posit.co/>
5. Hirukawa, M., Prokhorov, A.: Consistent estimation of linear regression models using matched data. *J. Econometrics* **203**(2), 344–358 (2018). <https://doi.org/10.1016/j.jeconom.2017.07.006>
6. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, vol. 73. Wiley, New York (2019)
7. Moretti, A., Shlomo, N.: Improving statistical matching when auxiliary information is available. *J. Surv. Stat. Methodol.* **11**(3), 619–642 (2023). <https://doi.org/10.1093/jssam/smac038>
8. Okner, B.: Constructing a new data base from existing microdata sets: the 1966 merge file. *Ann. Econ. Soc. Meas.* **1**(3), 325–362 (1972)
9. Rubin, D.B.: Characterizing the estimation of parameters in incomplete-data problems. *J. Am. Stat. Assoc.* **69**(346), 467–474 (1974). <https://doi.org/10.1080/01621459.1974.10482976>
10. Rubin, D.B.: Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.* **4**(1), 87–94 (1986). <https://doi.org/10.1080/07350015.1986.10509497>
11. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: On methods of statistical matching with and without auxiliary information. Technical report (1990)
12. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Surv. Pract.* **19**(1), 59–79 (1993)



On the Use of Simulations in the Construction of Composite Indexes

Giulio Grossi and Emilia Rocco^(✉)

Department of Statistics, Computer Science, Applications, University of Florence,
Firenze, Italy
`{giulio.grossi,emilia.rocco}@unifi.it`

Abstract. The interest in summarizing complex and multidimensional phenomena such as development, sustainability, poverty, quality of life, and well-being for broader comprehension, even among non-experts, is far from waning. Composite indexes stand out as the most commonly utilized methodological tool for this purpose. Various composite indexes have been proposed in the literature, often driven by the need to streamline a particular phenomenon. Each index is built upon specific assumptions, and its robustness for a given application is often evaluated afterwards, through uncertainty and sensitivity analyses. The performance of an index relies on the nature and characteristics of the indicators it aggregates, as well as on the relationships among them. We address these questions by investigating, through a simulation study, two distinct approaches to construct a composite index. The objective is to identify elements that, considering the available data and their characteristics, enable the selection of the most appropriate synthesis rule whatever is the multidimensional phenomenon under investigation.

Keywords: Autoencoder · cross-testing · principal component analysis · sensitivity to misspecification

1 Introduction

Composite indexes, statistical measures that combine a set of elementary (or individual) indexes and/or variables into a single measure, represent the main methodological approach used to describe complex and multidimensional phenomena. Proposed in literature across different fields of application, most of them rely on averages of elementary indices. These span from the simple arithmetic mean to more sophisticated formulas, potentially including weighted and penalized components. A comprehensive review of these methods can be found in [1], which offers valuable insights into the most commonly employed methodologies. Another research direction focuses on creating synthetic indexes as a dimensionality reduction task. In this approach, the primary objective is to condense the most pertinent information from a collection of elementary indicators into the synthetic index. Principal Component Analysis (PCA) is commonly employed

in this context, yet it is constrained by its ability to capture solely linear relationships through orthogonal projections. To address this limitation, a novel approach utilizes autoencoders [2] for dimensionality reduction. This method offers greater flexibility than PCA by capturing both linear and non-linear relationships between variables, leading to a more comprehensive understanding of the phenomenon.

To the best of our knowledge, there is no study in the literature that compares at least some of the several existing composite indexes. Typically, a particular composite index is recommended for analyzing a specific phenomenon. In evaluating the results obtained, such as ranking geographical areas, a sensitivity analysis is often conducted. This analysis may include assessing the synthesis method, as well as any weight systems or normalization forms applied to the elementary indices. Alternatively, the coherence of results with other existing analyses of the same or related phenomena is examined. The objective of this study is to compare, through simulations, the two aforementioned approaches for constructing a composite index. Specifically, the study examines four indexes, two from each approach. In the subsequent section, a concise review of the four indexes is presented alongside the criteria employed for their comparison. Section 3 delves into the description of the simulation study and subsequent discussion of the results.

2 Methodology

To evaluate the information compression capabilities of composite indexes across a spectrum of scenarios, which differ in terms of the characteristics of the synthesized elementary indices and their relationships, we compare four indexes documented in the literature. These indexes can be classified into two categories. The first category comprises latent variables constructed by combining elementary indices, potentially using weights and/or penalties defined by experts. Conversely, the second category consists of latent variables derived from elementary indicators, employing dimensionality reduction techniques such as PCA or autoencoder-based methods.

In the first category, the most common method for constructing a composite index relies on the simple but effective approach of averaging elementary indices. However, this method faces limitations, particularly the compensatory issue that is the fact that with this approach opposing trends in individual indices can cancel each other out in the synthetic measure, obscuring their distinct characteristics. To address this challenge, researchers have developed non-compensatory indices that avoid canceling out opposing trends. One such example is the Adjusted Mazziotta-Pareto Index (AMPI), the detailed implementation of which is presented in [3].

The composite indexes belonging to the second category are non-compensatory too, because they are derived by extracting information from the elementary indicators. In this sense, indexes of this type serve as a compressed representation of the entire data matrix. Techniques such as PCA, as

seen in [4], fall into this category, as well as the novel approach based on the use of autoencoders for dimensionality reduction [2]. Autoencoders offer greater flexibility compared to PCA, as they can capture nonlinear relationships beyond orthogonal projections, thus overcoming the limitation of PCA in capturing only linear relationships.

We aim to compare the synthesis capability of the four mentioned indexes from various perspectives. First of all, we want to evaluate the indexes' ability to reflect elementary variables. For this purpose, we use the concept of multidimensional stress measurement, which is defined as:

$$\theta = \sqrt{\frac{\sum_{i < j} (d_{ij} - \delta_{ij})^2}{\sum_{i < j} d_{ij}^2}} \quad (1)$$

with d_{ij} representing the euclidean distance between observations i and j in the original space, and δ_{ij} representing the same distance in the lower dimension space. Lower stress values indicate a composite index that more accurately represents the relationships and distances between the original data points. This suggests that the indicator captures the underlying data structure and patterns, faithfully reflecting the phenomenon of interest.

Another important characteristic of composite indexes to be evaluated, in our opinion, is their sensitivity to model misspecification. To achieve this objective, we have developed a suitable statistic, as described in the following. Let $R(X_i)$ represent the rank of observation i in the composite index X , and $\mathcal{R}(X)$ the vector of ranks for all observations. We compare these ranks in two scenarios: first, with full information available, and second, $R^{(k)}(X_i)$ estimated from a *undercomplete dataset*: e.g.: the original set of indicators minus the k -th indicator. We define a measure of sensitiveness to missing indicators in the dataset as

$$\rho = \frac{1}{K} \sum_{k=1}^K \frac{|R^{(k)}(X_i) - \mathcal{R}(X_i)|}{N} \quad (2)$$

3 Simulations

This section provides Monte Carlo simulation results to assess the performance of the four aforementioned composite indexes. We have considered $3 \times 3 \times 2$ scenarios derived from the combination of three distinct data generating processes (DGPs), three varying sample sizes, and two alternative approaches for the initial transformation of elementary indices. Regarding the data generation process, our objective was to explore how the composite indexes perform when they are obtained from elementary indices having disparate distributions which may also deviate significantly from normality assumptions, and to account simultaneously for the presence or absence of correlation between the elementary indices. In terms of sample size, we aimed to ascertain whether the increase of observation units had equivalent effects on the various composite indices. Lastly, concerning the initial transformation of elementary indices - aimed at rendering

them comparable for fair aggregation - we focused solely on comparing the use of a specific method against the option of directly operating on the individual indices in their original scales, we follow the normalization method described in [3], without exploring various normalization methods. The three utilized DGPs are as follows:

- **IID variables:** elementary indicators are independent and identically distributed, with Normal distribution
- **Correlated ID variables:** elementary indicators are correlated, but all of them follow the same distribution, and also in this case, we opt for the normal distribution. Correlation between elementary indicators range between -0.87 and 0.79 .
- **Correlated, no ID distributions:** elementary indicators are correlated and follow different distributions: 2 uniform distributions, a χ^2 distribution, a Poisson distribution, an exponential and a student's T. The correlation across elementary indicators range between -0.85 and 0.91 .

For each of the 9 scenarios, we have considered 100 replication. The values obtained for each of the four used aggregation methods are, firstly compared in terms of fit, measured by the stress index presented in Eq. 1.

Table 1 and Fig. 1 report the average stress test and the distribution of stress values in the different scenarios. Stress values are interpreted with respect to the original dataset in both cases, normed and non-normed indicators.

Based on these findings, it can be concluded that aggregation methods employing autoencoders generally yield better results compared to other methods. Particularly noteworthy is the observation that stress decreases for this method as sample size increases under non-normed scenarios. Conversely, the other three methods appear to be minimally affected by sample size, exhibiting similar performances across the different scenarios. Additionally, lower stress values are attained when dealing with correlated outcomes for both the two indexes derived from dimensionality reduction methods and the two calculated using aggregation methods such as arithmetic mean and AMPI.

Looking closely at the results derived from the two synthesis methods based on information reduction in scenarios where elementary indices have been normalized, we observe that the stress value for both the two composite indexes exceeds the outcome obtained without normalization. While this outcome might seem unexpected, it is logical: normalization techniques are crucial for ensuring comparability among elementary indices in weighted average methods. However, they alter the original distribution of indicators by compressing observations toward the distribution's center, thereby altering the relative distances between units.

The results concerning the "sensitivity" evaluation of the four aggregation methods, measured through the index introduced in Eq. 2, are shown in Fig. 2. For this analysis, we focus only on the 6 scenarios in which DGPs, are the last two: the ID, correlated scenario (first row) and the Non ID, correlated case (second row).

Table 1. Average stress values for Autosynth and alternative methods, over three DGP scenarios, 3 different sample sizes and normalization choices

	IID				Correlated - ID				Correlated - no ID			
	Autosynth	PCA	AMPI	Mean	Autosynth	PCA	AMPI	Mean	Autosynth	PCA	AMPI	Mean
Non-normed, N = 50	0.607	0.891	0.956	0.85	0.523	0.806	0.844	0.861	0.535	0.550	0.790	0.782
Non-normed, N = 250	0.581	0.878	0.843	0.84	0.418	0.746	0.815	0.836	0.406	0.754	0.790	0.785
Non-normed, N = 1000	0.547	0.810	0.840	0.84	0.401	0.870	0.820	0.841	0.394	0.723	0.794	0.781
Normed, N = 50	0.822	0.941	0.956	0.85	0.791	0.924	0.939	0.861	0.740	0.789	0.901	0.782
Normed, N = 250	0.834	0.958	0.959	0.84	0.818	0.953	0.945	0.836	0.725	0.849	0.912	0.785
Normed, N = 1000	0.837	0.961	0.963	0.84	0.824	0.955	0.951	0.841	0.778	0.873	0.927	0.781

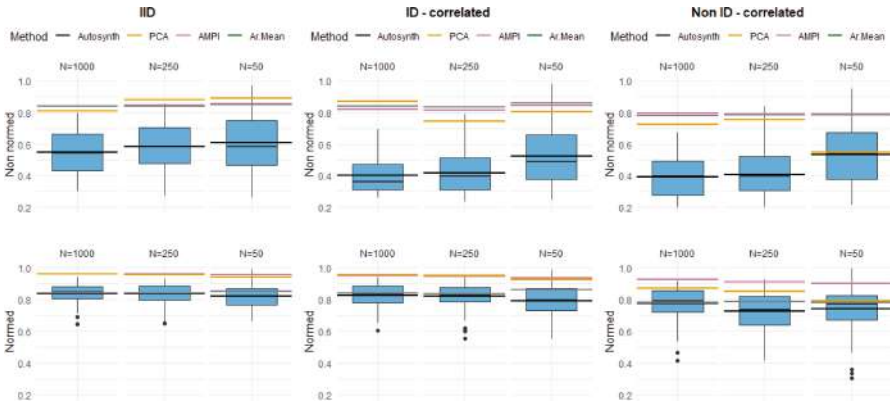


Fig. 1. Stress values for Autosynth and alternative methods, over three DGP scenarios, 3 different sample sizes and normalization choices

We observe that, overall, methods relying on the aggregation of elementary indicators exhibit lower sensitivity compared to methods based on dimensionality reduction. This observation is especially significant when considering the latter case, wherein elementary indicators can display diverse distribution shapes. In general, we notice that the sensitivity of methods based on dimensionality reduction diminishes as the sample size increases, while the sensitivity of the methods based on aggregation of elementary indexes does not change with the sample size.

4 Conclusions

In this work, we have investigated the properties of four different methods to construct composite indexes, under several scenarios. By comparing the results in terms of the fit of the index and in terms of sensitivity, we find out that the methods based on autoencoders grant better results in terms of fit, but at the same time can be sensible to the omission of variables in the theoretical model. On the other hand, indexes generated using arithmetic mean or AMPI are less

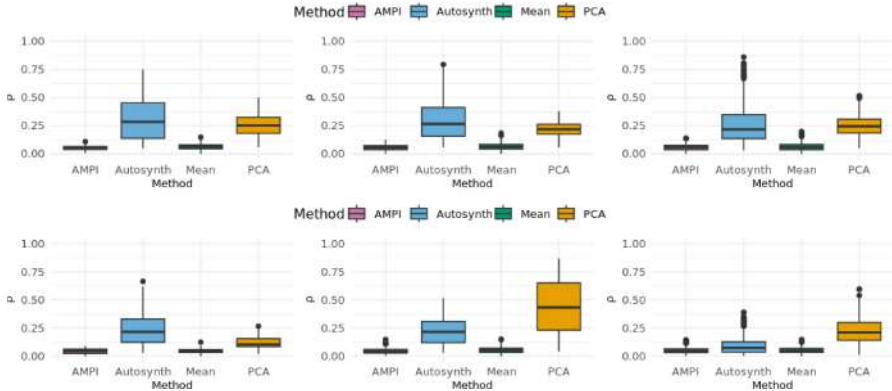


Fig. 2. Sensitivity to misspecification of the model for the alternative method. First row: ID and correlated elementary indicators, second row: No ID and correlated elementary indicators

sensitive but grant poorer fits. In overall, it is reasonable to think that methods that align better with the observed data are more sensible to variable omission, and vice versa. The practitioners should keep in mind what is their objective, whether to represent at best the phenomenon or a robust index even if it fits better the phenomenon. We also emphasize the significance of choices related to normalization, underlying distributions, and sample sizes, aspects that are not always explicitly addressed in existing literature on the topic.

References

1. Greco, S., Ishizaka, A., Tasiou, M., Torrìsi, G.: On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. *Soc. Indic. Res.* **141**, 61–94 (2019)
2. Grossi, G., Rocco, E.: AutoSynth index: a synthetic indicator for socio-economic development based on autoencoders. In: *Cladag 2023 Book of Abstracts and Short Papers*, pp. 507–510. Pearson (2023)
3. Mazziotta, M., Pareto, A.: Measuring well-being over time: the adjusted Mazziotta-Pareto index versus other non-compensatory indices. *Soc. Indic. Res.* **136**, 967–976 (2018)
4. Mishra, S.K.: A comparative study of various inclusive indices and the index constructed by the principal components analysis (2007). Available at SSRN 990831



Synthetic Control Method for Clinical Trials and Precision Medicine

Veronica Ballerini and Giulio Grossi^(✉)

Department of Statistics, Computer Science and Applications, University of Florence,
50144 Florence, Italy
giulio.grossi@unifi.it

Abstract. The large data availability fuels research on “one patient, one treatment” personalized medicine methods in clinical studies, particularly impactful for rare diseases where randomized controlled trials are impractical. Existing approaches like “external controls” using data from similar patients can be problematic in cases where there are limited suitable controls. This work proposes a novel approach combining historical controls with an adaptive study design, allowing treatment assignment to adapt based on valid estimates of treatment effects obtained within patient subpopulations. The proposed method represents a novel use of the “econometric” synthetic control method within clinical studies. We use a Bayesian approach to inference, which makes it easy to quantify the uncertainty related to the estimates and propagate it to the next patient’s treatment assignment. The novel combination of synthetic controls with adaptive designs leverages the flexibility of both methods for the effective handling of complex data patterns, improving treatment allocation efficiency while guaranteeing ethical studies. Simulation results support the promise of this approach for clinical studies.

Keywords: adaptive designs · synthetic control method · adaptive sampling · staggered adoption

1 Introduction

In recent years we are witnessing a paradigm shift in the development of new drugs and therapies; from the “one drug, one target disease”, to the “one patient, one treatment” strategy. This is primarily pushed by technological and medical advances, thanks to which we are increasingly able to identify predictive biomarkers. As the paradigm changes, the way new drugs or new medical therapies are tested must also change: in fact, classical randomized controlled trials (RCTs) are inefficient in this context. Going beyond RCTs, a notable trend is the utilization of “external controls,” which leverage data on similar patients from past trials or other research sites [5]. While appealing for studying treatment

Supported by the European Union - Next GenerationEU, UNIFI Young Independent Researchers Call - BayesMeCOS Grant no. B008-P00634.

effects without randomization, using external controls can be problematic in the context of rare diseases due to the limited availability of appropriate controls with comparable characteristics. To overcome this issue, we propose a strategy to fuse historical controls to a new “synthetic” one, which is directly comparable with the treated unit(s). To further improve the study’s efficiency and ethics, we propose to combine the synthetic control strategy with an adaptive design. Such an approach allows for the adjustment of the probability of assigning a new investigational treatment based on the so-far estimated effects within subpopulations of patients. We expect this strategy to lead to a more efficient allocation of treatments, particularly for rare diseases with limited patient availability. Our work boasts several innovative aspects. First, alongside independent research by [4], it represents the first application of the panel data methods in the family of synthetic control methods within the realm of clinical studies. It is crucial to distinguish that the term “synthetic control” can encompass various uses of historical controls in the broader clinical literature. In this context, we specifically refer to the method developed by [1, 2].

2 Method

Consider the following experiment. Patients i , $i = 1, 2, \dots, N_1$ are enrolled in a study to test the efficacy of an innovative treatment. Each patient i may be assigned to the innovative treatment or the standard of care (SOC). Let Z_i denote the treatment assigned to the i^{th} patient. For the sake of simplicity and without loss of generality, let us assume the treatment assignment to be binary; the realization z of Z_i can be either 1 (if assigned to the innovative treatment) or 0 (if assigned to the SOC). Moreover, we collect information on N_0 units $j = 1, \dots, N_0$, from an offline dataset. Notice that this offline dataset can comprise information coming from external sources, historical data and previous experiments. Units j are units assigned to the SOC at a certain point in time; Z_j can only be equal to 0. All units i and j are observed through time, both before the treatment assignment and after. We denote with negative values the periods before the treatment assignment, i.e., the pre-treatment or $\{T^-\}$, where the outcome of interest and covariates are observed in the absence of intervention; with 0 the assignment time, and with positive values the periods after the assignment, i.e., the post-treatment or $\{T^+\}$. Therefore, $t = \dots, -2, -1, 0, 1, 2, \dots$, or $t = \{T^-, 0, T^+\}$. Notice that the pre- and post-treatment may have different lengths for each patient i or j ; we denote with t_i and t_j patient-specific timelines. Hence, units i may be assigned to either the innovative treatment or the SOC at time $t_i = 0$, whereas units j were only assigned to the SOC at time $t_j = 0$.

We leverage the pre-treatment information of patients j to identify clusters of observations according to a set of available covariates (e.g., biomarkers, gender, $\dots$), and pre-register the subpopulations. We denote the baseline covariates with the vector $\mathbf{X}$. Assume to identify G subpopulations, and let g be the subpopulation index, $g = 1, \dots, G$. At the beginning of the experiment, I_0 , let the probability for units in subpopulation g to be assigned treatment $Z_i = 1$ be

equal to $p_{0,g} = p_0$. Assume to perform an interim analysis in I_1 to assess the heterogeneous efficacy of the new treatment on an outcome of interest, Y . It would be convenient, more efficient, and ethical to assign to the next enrolled patient belonging to subpopulation g the treatment that has been found to be more beneficial for that subpopulation. Hence, let $p_{1,g}$ be an “updated” probability to be assigned to the innovative treatment for patients belonging to g that are enrolled in the study starting from I_1 . Such probability $p_{1,g}$ deviates from $p_{0,g}$ in that it takes into account the estimated treatment effect for units in g .

We frame our analysis under the Rubin Causal Model. Under SUTVA (Stable Unit Treatment Value Assumption), let Y_{i,t_i}^- be the observed outcome of the i^{th} unit at time $t_i \in T^-$, and similarly for Y_{j,t_j}^- . Let $Y_{i,t_i}(z)$ be the potential outcome for the i^{th} unit at time $t_i \in T^+$ under treatment z . For units j , we can only define $Y_{j,t_j}(0)$. Overall, we can define the observed outcomes as follows.

$$\begin{aligned} Y_{i,t_i} &= \mathbb{I}_{\{t_i \in T^-\}} Y_{i,t_i}^- + \mathbb{I}_{\{t_i \in T^+\}} [Z_i Y_{i,t_i}(1) + (1 - Z_i) Y_{i,t_i}(0)] \\ Y_{j,t_j} &= \mathbb{I}_{\{t_j \in T^-\}} Y_{j,t_j}^- + \mathbb{I}_{\{t_j \in T^+\}} Y_{j,t_j}(0) \end{aligned} \quad (1)$$

We identify the treatment effect on the treated units as follows:

$$\tau_{i,t_i} = Y_{i,t_i}(1) - Y_{i,t_i}(0) \quad \forall i : Z_i = 1, t_i \in \{T^+\} \quad (2)$$

In clinical trials, such estimand can be of great interest. As highlighted by [4], focusing on the ATE can lead to disregarding the effect of treatment in small subpopulations. Precision medicine instead aims to identify such subpopulations and estimate their causal effects.

Following [1, 2], we impute the missing potential outcomes $Y_{i,t_i}(0)$ of the units i assigned to the innovative treatment as a linear combination of time series for the external control units, as follows

$$Y_{i,t_i}(0) = \sum_{j=1}^{N_0} \omega_j Y_{j,t_j} \quad \forall i : Z_i = 1, t_i = t_j \quad (3)$$

where the weights ω_j are derived as solution of a vertical regression minimization problem of the form

$$\begin{aligned} \tau_{i,t} &\sim \mathcal{N} \left(Y_{i,t} - \sum_{j=1}^{N_1} Y_{j,t} \omega_j, \sigma^2 \mathbf{I} \right) \\ \omega &\sim \mathcal{N}(0, 0.1), \sigma^2 \sim \Gamma^{-1}(1, 1) \end{aligned} \quad (4)$$

Once synthetic control weights are estimated, we calculate the posterior distribution for the missing outcome $Y_{i,t_i}(0)^*$ and derive the posterior distribution for the ATT (Average treatment effect on the treated) : $\tau_{i,t_i}^* = Y_{i,t_i} - Y_{i,t_i}(0)^*$. We opt for a Bayesian method due to its strength in uncertainty quantification. Unlike traditional methods within vertical regression and synthetic control contexts, it provides a clear posterior distribution for the treatment effect, aiding

interpretation. Furthermore, once the posterior distribution $Y_{i,t_i}(0)^*$ is derived, obtaining any function of it is straightforward. Let $\{P_h\}$ be a Markov Process, with $P_h \in (0, 1)^G$, where the G elements of P_h are independent; $h = 1, 2, \dots$, denotes the interim analysis conducted to evaluate the effect of the treatment. Let also each element $p_{h,g}$ of P_h follow a Beta distribution; according to [3] parametrization,

$$p_{h,g} \sim B(\mu_{h,g}, \phi_g) \quad \forall t \quad (5)$$

where $\mu_{h,g} = p_{h-1,g}(1 + \beta\tilde{\tau}_{h,g})$, and ϕ_g is a cluster-specific precision parameter. The variable $\tilde{\tau}_{h,g}$ is the posterior relative amplitude of the treatment effect with respect to the outcome in the absence of treatment, estimated aggregating the effects of all units of cluster g treated up until h , i.e., $\tilde{\tau}_{h,g} = \{\tau_{i,t_i}^*/Y_{i,t_i}(0)^* : i \in g, i \text{ is assigned to treatment before } h\}$. The sensitivity parameter β represents the degree of adaptiveness of the trial; e.g., $\beta = 1$ means that a 10% posterior treatment effect-outcome ratio would cause an equivalent variation in assignment probability. The location parameter $\mu_{h,g}$ indicates that the probability of assigning the new treatment after the h -th analysis to a patient of cluster g is centred on the probability at $h - 1$ plus a function of the treatment effect. We use the described design in the simulation in Sect. 3; however, any further specifications of $\mu_{h,g}$ are allowed.

3 Simulation

Our simulation replicates a (simplified) Oncological clinical trial where $N_1 = 100$ enrolled patients receive the same initial standard preoperative chemotherapy and surgical tumour resection. Then, each patient is assigned to either standard ($Z = 0$) or innovative ($Z = 1$) postoperative therapy. In this simulation, we arbitrarily assume all patients are observed for about 8 months (35 weeks; $|T^-| = 35$) before the surgery, and for 15 weeks after the resection ($|T^+| = 15$). Historical data on $N_0 = 200$ patients receiving standard postoperative therapy are available from other hospitals' databases; we assume that such data contain similar longitudinal information. Our outcome of interest is a continuous variable, $Y \in \mathbb{R}^+$, indicating the severity of the patient's status. We assume $Y > 200$ to indicate a severe condition requiring surgery, and we simulate data accordingly. We also assume that, based on historical data, it is possible to identify $G = 4$ clusters of patients according to their prognostic biomarkers, and that it is easy to assign each patient the proper cluster. Without loss of generality, potential outcomes are simulated such that the innovative postoperative therapy is: very effective and effective (negative ATT) for patients of clusters 1 and 2, respectively; not significant for patients of cluster 3 (but with slightly positive ATT); harmful to patients of cluster 4 (positive ATT). To show the potentiality of the proposed method, we partition the sample unequally; each patient has a probability of 0.15 to be assigned to clusters 1 or 4, and 0.35 to be assigned to clusters 2 or 3. In our simulation patients are enrolled at any time, but, for the sake of simplicity and still without loss of generality, they are assigned to treatment once the previous patient finished the observation period. Hence, h

corresponds to the number of patients enrolled in the study until h minus one; e.g., the first patient of cluster g is enrolled in the study at $h = 0$ and assigned to the treatment with probability $p_{h=0,g} = p_0$. Specifically, in each cluster the first patient is assigned to the innovative treatment with probability $p_0 = 0.5$; then, $p_{h,g}$ is drawn according to Eq. (5), where $\beta = 1$. Figure 1 shows how $\mathbf{p}$ evolves as patients are enrolled in the study. For the smaller clusters, the probability of being assigned to the new treatment moves with caution; $p_{h,1}$ slowly increases up to 0.875 at $h = 6$, whereas $p_{h,4}$ decreases until 0.12 for the same h . Overall, almost 55% of patients belonging to cluster 1 are treated with innovative treatment; the proportion for cluster 4 is much lower, equal to a third. Clusters 2 and 3 are the larger clusters. Despite its appearance in Fig. 1, $p_{h,2}$ is around $p_{h,1}$ at $h = 6$; then, as the treatment proves to be effective for patients in cluster 2, it continues growing stabilizing above 0.5. However, despite its sharp increase, the distribution of $p_{h,2}$ keeps propagating a significant amount of uncertainty so that while the mean is set around 0.96 at the end of the experiment, its lower bound is slightly greater than 0.6. Concerning cluster 3, we remind that the simulated effect is slightly positive (harmful) but not significant; $p_{h,3}$ slowly decreases reaching 0.28 for $h = 12$; the intervals greatly cover 0.5. At the end of the experiment, the 36% of patients of cluster 3 are treated. As a major pro of our design, we may say that is shown to be quite conservative.

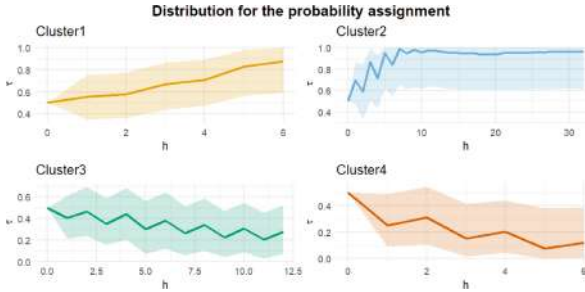


Fig. 1. Adaptive probability within clusters - shaded area: 95% BCI

Our simulation results, presented in Fig. 2, support the value of our proposed estimation strategy for accurately estimating causal effects on target subpopulations in adaptive trials. Notably, the estimated effects closely track the true effect trajectory, particularly for clusters 2 and 3, which have larger sample sizes.

4 Further Research

In this short paper, we proposed an ethical strategy to treat patients while assessing the effect of a new investigational treatment with respect to the standard of care. Simulation results underline promising results for the use of this strategy and class of estimators within clinical studies with panel data, with a limited

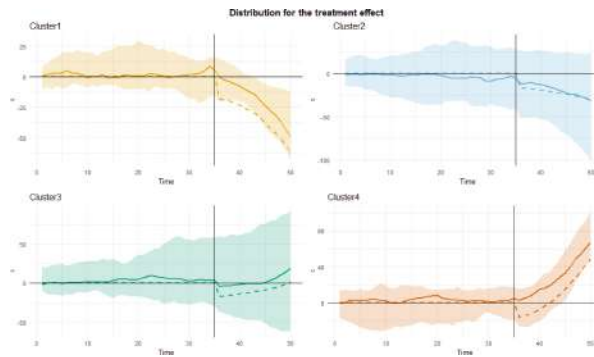


Fig. 2. Total ATT within clusters - shaded area: 95% HPD, solid line: synthetic control + adaptive design estimates, dashed line: true values

amount of resources needed. This preliminary work needs further research to address several relevant open questions. Among them are whether it is possible to think of a strategy to set β according to a Bayesian decision theory framework, and what the necessary sample size is to outperform an RCT in terms of credibility intervals of the estimated treatment effect.

References

1. Abadie, A., Diamond, A., Hainmueller, J.: Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J. Am. Stat. Assoc.* **105**(490), 493–505 (2010)
2. Doudchenko, N., Imbens, G.W.: Balancing, regression, difference-in-differences and synthetic control methods: a synthesis. Technical report, National Bureau of Economic Research (2016)
3. Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. *J. Appl. Stat.* **31**(7), 799–815 (2004)
4. Hüyük, A., Qian, Z., van der Schaar, M.: Adaptive experiment design with synthetic controls. arXiv preprint [arXiv:2401.17205](https://arxiv.org/abs/2401.17205) (2024)
5. Thorlund, K., Dron, L., Park, J.J.H., Mills, E.J.: Synthetic and external controls in clinical trials—a primer for researchers. *Clin. Epidemiol.* 457–467 (2020)



Functional Clustering with Imbalanced Groups

Catherine Higgins^(✉) and Michelle Carey

School of Mathematics and Statistics, University College Dublin, Dublin, Ireland
catherine.higgins1@ucdconnect.ie, michelle.carey@ucd.ie

Abstract. Class imbalance, where some clusters have significantly more curves than others, pose difficulties for functional clustering methods. Often, the largest clusters are prioritized at the expense of smaller clusters, resulting in a misclassification of curves. We propose functional iterative hierarchical clustering as a novel approach to cluster functional data with imbalanced classes. We compare our proposed method to existing functional clustering methods through a simulation study and assess the performance on a benchmark dataset. The proposed approach demonstrates improved clustering accuracy in terms of the adjusted rand index, achieving on average a 22% improvement over existing methods.

Keywords: Functional Data Analysis · Clustering · Class Imbalance, · Hierarchical Clustering

1 Introduction

Functional clustering techniques classify a sample of curves into K groups of curves, without prior knowledge of the underlying clustering structure. The class imbalance problem, that is where the number of curves in one cluster is considerably more than the number of curves in another cluster, is a challenging issue. Some minor clusters, clusters with a small number of curves, are often ignored and/or misclassified into the major clusters with large numbers of curves, which leads to a poor clustering accuracy and a distortion of the true underlying cluster structure.

Imbalanced classes occur in numerous real-world situations, such as detecting faults in industrial machinery and identifying genes affected by infection or disease. While the class imbalance problem has been extensively studied in the context of supervised classification [1, 2], it has received limited attention in an unsupervised environment. Specifically, the method proposed in [3] stands as the sole functional clustering approach addressing this issue.

We propose functional iterative hierarchical clustering as a novel approach for clustering functional data with imbalanced classes. We compare to existing functional clustering methods namely, FADP_fPCs and FADP_L2 [3], funFEM [4], funHDDC [5], EM_fPCs [6], kCFC_fPCs [7] and funclust [8]. We then assess the performance on the Berkeley Growth Study benchmark dataset [9].

2 Functional Iterative Hierarchical Clustering

Consider a set of N curves that evolve over time $x_1(t), \dots, x_N(t)$, where $t \in [0, T]$. Our aim is to cluster these curves into K groups. The observed values of the sample curves usually contain observational errors, that is $y_{i,j} = x_i(t_{i,j}) + \epsilon_{i,j}$, for $i = 1, \dots, N$ where i represents the curves and $j = 1, \dots, M_i$ where j indexes the sampled time points within each curve, $t_{i,j}$ denotes the j^{th} time at which we sample the i^{th} curve and the errors $\epsilon_{i,j}$ are assumed to be independently distributed with mean 0 and a constant variance σ^2 . The Functional Iterative Hierarchical Clustering (funIHC) approach carries out the following steps:

1. **Reconstruct the functional form of the data:** Given the observed values $y_{i,j}$ for $i = 1, \dots, N$ and $j = 1, \dots, M_i$, the functional form is reconstructed from the discrete observations by approximating the curves via a basis function expansion, that is, $\hat{x}_i(t) = \sum_{k=1}^K c_k \phi_k(t)$, where $\phi_k(t)$ represents basis functions, commonly B-spline basis functions, and c_k are the associated coefficients. To estimate the coefficients of the basis function expansion $\{c_k : k = 1, \dots, K\}$ use the standard penalized least squares smoothing approach [10].
2. **Clustering procedure:** Let the $N \times M$ matrix $\mathbf{A}$ be the estimated first derivative of the curves evaluated at an equally spaced grid of points over the domain. The rows of $\mathbf{A}$ represent individual curves and columns denote the estimated values for these curves. Let α_{min} and α_{max} be the minimum and maximum of the Spearman rank correlation between all the possible pairs of the rows of $\mathbf{A}$. Define $[\alpha_{min}, \dots, \alpha_{max}]$ as a grid of Q equally spaced values between α_{min} and α_{max} . Cluster the rows of $\mathbf{A}$ using the iterative hierarchical clustering method as described in [11], where the Spearman rank correlation is the distance metric with a threshold of $1 - \alpha_q$, where q ranges from 1 to Q . The optimal value α_{opt} is obtained by maximizing the average silhouette value, which indicates how closely a curve aligns with other curves in its own cluster compared to those in different clusters. The value ranges from -1 to 1, where higher values indicate greater similarity with curves in the assigned cluster and lower similarity with curves in other clusters.

3 Simulations

The simulated sample curves, X_i , are realisations of a Gaussian process with Matérn covariance function $C(t, t') = 0.2 \times \exp(-0.3\|t - t'\|)$ over a discrete grid of domain $I = [0, M]$ [12]. We define six different mean functions: $\sin(2\pi t)$, $\cos(2\pi t)$, $\sin(4\pi t + \pi/2)$, $\sin(4\pi t - \pi/2)$, $\sin(3\pi t + \pi/3)$ and $\sin(6\pi t - \pi)$ and six different cluster sizes, that is $N_1 = 500$, $N_2 = 500$, $N_3 = 10$, $N_4 = 15$, $N_5 = 3$, and $N_6 = 200$. A sample of curves of size N is obtained by $Y_{i,j} = X_i(t_{i,j}) + \epsilon_{i,j}$ for $i = 1, \dots, N$ and $j = 1, \dots, M$, where $\epsilon_{i,j}$ is a normally distributed random variable with mean 0 and standard deviation σ_ϵ . The simulation study includes

two distinct sample sizes: a small sample size comprising $M = 15$ time points and a larger sample size with $M = 200$ time points. Furthermore, the study accounts for two levels of error: $\sigma_\epsilon = 0.05$ and $\sigma_\epsilon = 0.15$.

We compare funIHC with the following seven state-of-the-art functional clustering methods: FADP_fPCs and FADP_L2 [3], funFEM [4], funHDDC [5], EM_fPCs [6], kCFC_fPCs [7] and funclust [8]. Table 1 presents the average Adjusted Rand index [13] (μ_{ARI}) and the Davies-Bouldin index [14] (μ_{DB}) for each of the functional clustering methods. The Adjusted Rand index ranges from -1 to 1, where 0 indicates random clustering and 1 indicates identical clustering relative to the true clustering partition. The Davies-Bouldin index measures the within-cluster distances to between-cluster distances, with the lowest value indicating the best clustering partition. funIHC demonstrates improved performance over competing methods, achieving the highest μ_{ARI} and the lowest μ_{DB} for the four different simulated scenarios.

Table 1. The average Adjusted Rand Index (μ_{ARI}) and the Davies-Bouldin index (μ_{DB}) for the eight functional clustering methods across four different simulated scenarios.

Method	M = 15		M = 15		M = 200		M = 200	
	$\sigma_\epsilon = 0.05$		$\sigma_\epsilon = 0.15$		$\sigma_\epsilon = 0.05$		$\sigma_\epsilon = 0.15$	
	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}
funIHC	1.00	0.80	1.00	0.86	1.00	0.82	1.00	0.87
EM_fPCs	0.98	0.91	0.98	0.99	0.98	0.86	0.98	0.97
FADP_fPCs	0.98	0.81	0.98	0.87	0.99	0.82	0.95	0.90
FADP_L2	0.97	0.85	0.96	0.97	0.95	0.84	0.95	0.90
kCFC_fPCs	0.92	0.96	0.86	1.07	0.68	1.11	0.68	1.20
funFEM	0.91	0.93	0.89	0.99	0.50	1.04	0.49	1.23
funclust	0.78	0.83	0.68	1.15	0.68	1.11	-0.005	35.9
funHDDC	0.54	0.94	0.50	0.97	0.48	0.85	0.47	0.95

4 Benchmark Data Berkeley Growth Study

Originating from the Berkeley growth study [9], this dataset is available in the R package `fda`. In this dataset, the heights of 54 girls and 39 boys were measured at 31 stages, from 1 to 18 years. Figure 1 illustrates the individual growth curves in grey and the centre of the cluster obtained by funIHC in black. The centres of the clusters for the growth velocity curves (rate of change of height over time) are also displayed together to aid interpretation.

funIHC clustered the growth velocity curves into three groups. The first cluster contains 54 children (4 males and 50 females) who display a peak in growth

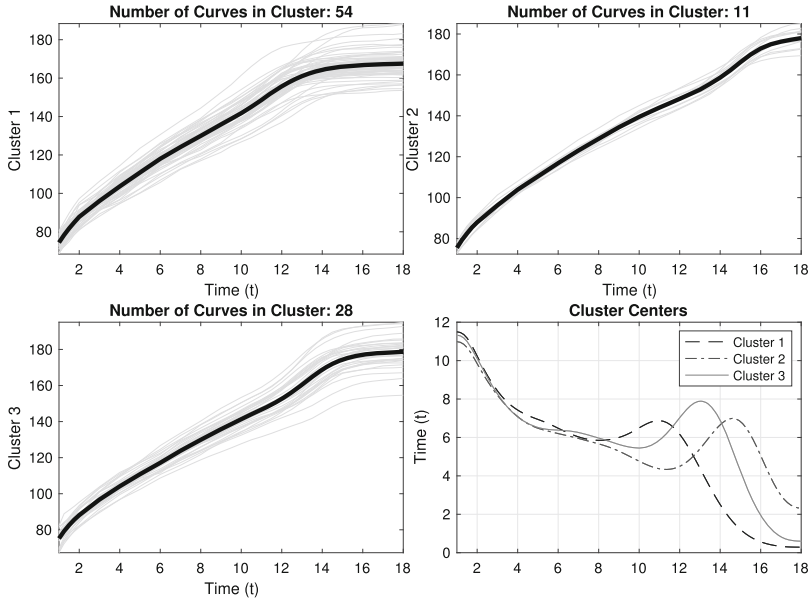


Fig. 1. Clustering obtained by funIHC for the Berkeley Growth dataset. Individual growth curves are in grey and the centre of the cluster in black, along with the centres of the velocity of the clusters.

velocity around 11 years of age. The second cluster contains 11 children (all male) who display a sharper decline in the growth velocity from 5–11, with a delayed peak in growth velocity around 14.5 years of age. The third cluster contains 28 children (24 males and 4 females) who display a higher peak in growth velocity around 13 years of age.

5 Conclusion

Functional Iterative Hierarchical Clustering (funIHC) provides a novel approach for clustering functional data with imbalanced classes ensuring that there is an adequate separation between different clusters and a sufficient homogeneity within any given cluster. As demonstrated through the simulation study, funIHC offers improvement, in terms of the adjusted rand index, of on average 22% over existing methods when clustering data that suffers from class imbalance issues. In addition, in each of the four simulated scenarios it achieves an Adjusted Rand index of 1, being the only approach to achieve the true clustering partition and the lowest Davies-Bouldin index, indicating it achieves the best clustering partition out of all methods. funIHC provides a meaningful clustering and interpretation of the Berkeley Growth data, with clusters distinguished based on the peak of the velocity observed in the growth curves. The R and MATLAB code for implementing funIHC is available at www.fdaatucd.com.

Acknowledgments. This research was funded by Science Foundation Ireland through the SFI Centre for Research Training in Geonomics Data Science under Grant number 18/CRT/6214.

References

1. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* **73**, 220–239 (2017)
2. He, H., Ma, Y.: *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley, New York (2013)
3. Ren, R., Fang, K., Zhang, Q., Wang, X.: Multivariate functional data clustering using adaptive density peak detection. *Stat. Med.* **42**(10), 1565–1582 (2023)
4. Bouveyron, C., Côme, E., Jacques, J.: The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.* **9**(4), 1726–1760 (2015)
5. Bouveyron, C., Jacques, J.: Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.* **5**(4), 281–300 (2011)
6. Chen, W.C., Maitra, R., Melnykov, V.: EMCluster: EM algorithm for model-based clustering of finite mixture Gaussian distribution. *R Package* (2012). <http://cran.r-project.org/package=EMCluster>
7. Chiou, J.-M., Li, P.-L.: Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B Stat Methodol.* **69**(4), 679–699 (2007)
8. Jacques, J., Preda, C.: Funclust: a curves clustering method using functional random variables density approximation. *Neurocomputing* **112**, 164–171 (2013)
9. Tuddenham, R.D., Snyder, M.M.: Physical growth of California boys and girls from birth to eighteen years. University of California publications in child development (1954)
10. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*, Springer, New York (2005)
11. Carey, M., Wu, S., Gan, G., Wu, H.: Correlation-based iterative clustering methods for time course data: the identification of temporal gene response modules for influenza infection in humans. *Infect. Disease Model.* **1**(1), 28–39 (2016)
12. Matérn, B.: *Spatial Variation*, 2nd edn. Springer, New York (1986)
13. Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* **2**, 193–218 (1985)
14. Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 224–227 (1979)



A Triad Census-Based Synthetic Indicator to Summarise Cooperation in Football

Riccardo Ievoli¹(✉), Lucio Palazzo², Roberto Rondinelli²,
and Giancarlo Ragozini²

¹ Department of Chemical, Pharmaceutical, and Agricultural Sciences,
University of Ferrara, Ferrara, Italy
riccardo.ievoli@unife.it

² Department of Political Sciences, University of Naples Federico II, Naples, Italy

Abstract. This work discusses the introduction of a synthetic indicator to summarise the triadic information in a football passing network taking also into account the number of passes occurring between players. To this aim, we aggregate the counts of triads by using a weighting scheme based on the number of connections in each isomorphism class. We compute this measure by using different cutoffs on the weighted network of passes between players. The final indicator can be obtained by averaging the triad-based indicators for a given number of cutoffs. The applicability of this approach is shown through a real dataset regarding three consecutive seasons of the UEFA Champions League.

Keywords: passing networks · triad census · synthetic indicator · network cutoffs

1 Introduction

In football analytics, the distribution of passes between players provides relevant information concerning the tactics and strategies. Since this distribution can be intended as a weighted and directed graph, network analysis (NA) has been applied to identify useful patterns regarding the different degrees of cooperation of individual players or the whole team. For instance, summary measures of passing networks have been used to model and/or predict football outcomes [1–3]. From a different perspective, NA can also be used to identify tactics and/or styles of play through cluster analysis [4].

Together with density, transitivity, or presence of clusters, network topological properties can be investigated through the triad census analysis [7]. In football passing networks, “triads” are defined as the interactions between three players. Given the small size of these kinds of networks, the count of triadic modalities appears well-suited to describe their topological structure. Sixteen triadic configurations (i.e., “isomorphism classes”), can be defined and their distribution is denoted as “triad census” [8]. In addition, the relevance of sequences of three passes in football is also pointed out in [6].

Furthermore, recent research proposes to test the triadic structure of football teams to assess the deviation from the randomness of the observed triad census [5]. The work considers three consecutive seasons of the UEFA Champions League (UCL) to study the observed distribution of triads in football [5]. Moreover, two possible issues arise. The first is that the triad census is suitable only for binary (unweighted) networks, neglecting the level of heterogeneity that can be observed in the passing distribution of players. Secondly, the construction of a unique “triadic” measure, characterizing the cooperation between the players of a team, represents an interesting challenge to understand different styles of play, especially from a ball possession perspective and/or the ability to involve many tactical solutions.

The present paper aims to propose a synthetic indicator based on the triadic census that can be used to characterize the ability of football teams to connect players. The proposed indicator is constructed as an aggregation of triad census-based measures obtained for different networks constructed by using incremental cutoffs starting from the original network. Each triad census-based measure is computed as a weighted average of the frequencies for the different isomorphism classes, considering a particular set of weights starting from zero under the full disconnected scenario and increasing with the number of links. The applicability of the synthetic indicator is shown through a real dataset including the matches of the Group Stage of the UEFA Champions League (UCL), with a specific focus on the comparison between the 14 teams involved in three consecutive seasons.

2 The Construction of a Triad Census-Based Indicator

The passing network of a football team is usually available in the form of an adjacency matrix and can be represented through a graph, see [2] (among others) for the technical details. In this contest, the minimal structural group in a network is defined as a “triad”. In directed networks there are 16 possible different triadic modalities, also denoted as *isomorphism classes*. Figure 1 summarizes these classes and their labels, from the so-called “003”, (zero edges between nodes), to the triad involving six links triads (“300”). Descriptive analysis regarding triad census in football can be observed in [5] (Table 4 and Fig. 4), also discussing the usefulness of triadic information to distinguish between football teams.

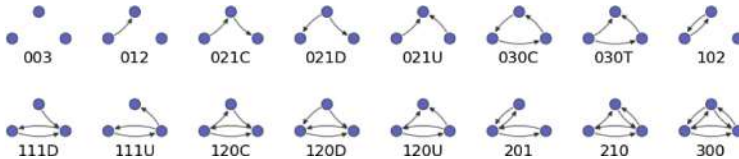


Fig. 1. List of the 16 isomorphism classes of the triad census.

For a given weighted and directed network $g = 1, \dots, G$ we define $\tau_{1,g}, \dots, \tau_{16,g}$ as the count of triads for each isomorphism class. For instance,

$\tau_{1,g}$ is the count of “003” triads while $\tau_{16,g}$ is the sum for the class “300”. We also define $g(0)$ as the binarized version of the network g , i.e., a network where players are connected if there is at least a pass between them. We remark that, using conventional algorithms, $\{\tau_{1,g}, \dots, \tau_{16,g}\} = \{\tau_{1,g(0)}, \dots, \tau_{16,g(0)}\}$. In addition, $g(c)$ is the binarized version of network g obtained by considering the c -th level (cutoff) in terms of the number of passes, with $c = 1, 2, \dots$. For instance, $g(2)$ presents edges between players only if there are at least three passes between them.

Given these premises, to construct a synthetic indicator for a football team in a match, we identify the following four main steps.

1. Obtain the triad census by using the binarized network $\tau_{1,g(c)}, \dots, \tau_{16,g(c)}$, starting from $c = 0$ (no cutoff).
2. Summarize the triadic information through a weighted average, denoted as $TC_{g(c)}$ by using a specific weighting scheme $w_1, \dots, w_{16}$.
3. Repeat steps 1–2 for c^* times where c^* is a given number denoting a specific number of passes.
4. Use a function to aggregate $TC_{g(0)}, TC_{g(1)}, \dots, TC_{g(c^*)}$ in a unique number.

After describing the general procedure, we back into the details of the steps. In particular, the indicator TC for the g -th network considering the cutoff c (Step 2) can be computed as follows:

$$TC_{g(c)} = \sum_{i=1}^{16} \tau_{i,g(c)} w_i \quad i = 1, \dots, 16. \quad (1)$$

In this context, the choice of w_i is non-trivial. For example, considering a fixed number of vertices (e.g., the 11 football players) and equal weights ($w_i = 1/16$) $TC_{g(c)}$, the total count of triads is always equal to $\binom{n}{3}$ for all $g = 1, \dots, G$. Then, we propose to consider weights that increase with the number of links in isomorphism classes (ranging from 0 to 6), associating a weight equal to zero to the disconnected triads and the maximum weight in the case of fully connected triads. This strategy allows rewarding the teams that present more connections (e.g., tactical solutions) in the different passing “levels” identified by cutoffs. The proposed system of weights is summarized in Table 1.

After selecting the number of cutoffs c^* , the aggregation of the $TC_{g(c)}$ in a unique index TC^* can be carried out through the arithmetic mean, the geometric mean, or the individual scores of a principal component analysis (PCA) among others.

3 Empirical Illustration

Data for the empirical analysis are the team passing distributions retrieved from the Group Stage of three consecutive seasons (2016–2017, 2017–2018 and 2018–2019) of the UCL¹. Then, a total of $G = 576$ passing networks coming from 288

¹ www.uefa.com.

Table 1. Weighting scheme to obtain $TC_g(c)$

Class/Classes	Number of Links	w_i
‘003’	0	0.0000
‘012’	1	0.0476
{‘021C’, ‘021D’, ‘021U’, ‘102’}	2	(0.0952)/4
{‘030C’, ‘030T’, ‘111D’, ‘111U’}	3	(0.1429)/4
{‘120C’, ‘120D’, ‘120U’, ‘120U’}	4	(0.1905)/4
‘210’	5	0.2381
‘300’	6	0.2857
	21	1

matches is considered (192 for each season). These data are discussed in detail in [5].

To guarantee complete comparability between teams, we use the passing networks considering the initial line-up ($v = 11$ players) such that each $g(c)$ network presents $\sum_{i=1}^{16} \tau_{ig} = 165$. For what concerns the choice of c^* we employ an empirical strategy based on the observed distribution of edges in the weighted networks. Then, the selected $c^* = 6$ is the overall median of the observed third quartiles of the networks’ weights. In addition, results of a PCA conducted on the seven triad-based sub-indicators ($TC_{g(0)}, \dots, TC_{g(6)}$) show that the first dimension accounts for the 88.1% of the entire variability. Aggregated measures TC^* obtained through arithmetic mean, geometric mean, and scores of the first principal component, exhibit a ranking correlation which is at least equal to 0.98. Then, we decided to analyze $TC_g^* = 7^{-1} \sum_{c=0}^6 TC_{g(c)}$. Considering descriptive statistics, $\overline{TC^*} = 11.7$ (standard deviation equal to 3.4), and its median accounts for 11.4 (Inter-quartile range equal to 4.6). The minimum is 3.8 and it is reached by the PSV Eindhoven team in its fourth match of season 2016–2017, while Manchester City exhibited the maximum (22.4) in season 2018–2019. Furthermore, the null hypothesis of normality cannot be rejected for the Anderson-Darling test considering a nominal level of 0.01.

Further analysis is carried out by focusing on the 252 networks (six per season) of the 14 teams that participated in all three considered seasons of the UCL: Atletico Madrid, Barcelona and Real Madrid (from Spain), Manchester City and Tottenham (England), Benfica and Porto (Portugal), Monaco and PSG (France), Bayern Munchen and Borussia Dortmund (Germany), Juventus and Napoli (Italy), CSKA Moskov (Russia). Figure 2a illustrates the distribution of TC^* for these teams. Considering the red dashed line as the overall median of the synthetic indicators, teams such as Monaco and CSKA Moskov perform worse in terms of cooperation for a large part of the matches, while Porto and Benfica appear close in terms of results. Atletico Madrid, Naples and Tottenham occupy an intermediate position, while teams such as Barcelona, Bayern Munchen and Real Madrid (presenting also a low variability) share similar results in terms of

high cooperation. Descriptive results are confirmed by using a conventional t-test for the difference in mean (Fig. 2b). For instance, considering Real Madrid, who won two editions over three (2016–2017 and 2017–2018), the null hypothesis of an equal mean of TC^* is not rejected in the comparison between Barcelona, Bayern Munchen, Borussia Dortmund and Manchester City. Conversely, regarding Porto, the null hypothesis cannot be rejected for CSKA Moskov, Benfica and Monaco.

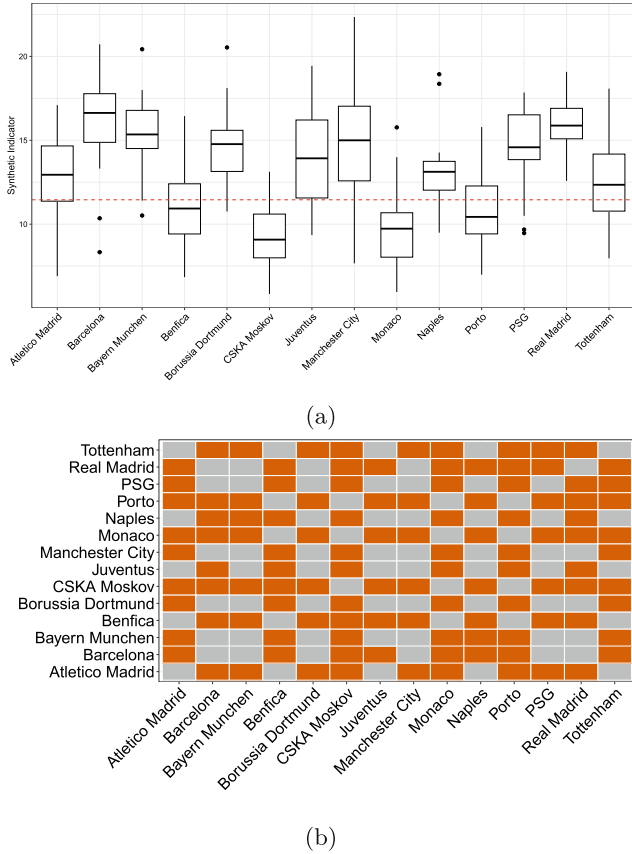


Fig. 2. Boxplot of the distribution of TC^* for 14 teams involved in UCL (panel 2a) and p-values of t-student two-sample test for the mean difference (orange cells depict p-values < 0.05).

4 Concluding Remarks

The present short paper shows the applicability of a new synthetic indicator of cooperation in football which is constructed starting from triad census analysis and different binary networks generated by the same distribution of passes.

Assuming a fixed number of vertices (e.g., 11 football players), this indicator can be computed without using a normalization technique given a choice of c^* , i.e., the number of layers (binary networks) that should be involved. Empirical results appear promising in terms of distinguishing between football teams in terms of cooperation. To summarize, the present approach represents a tentative to build a “bridge” between binary triad census and weighted NA.

Further research will regard an analysis of both the uncertainty and the sensitivity of the proposed synthetic indicators. In addition, a comparison between the indicator and other network summary measures at the team level (e.g., degree, centralizations and global clustering coefficients) will be carried out. The proposed approach will be also tested by considering networks of different sizes and different relational structures.

References

1. Grund, T.U.: Network structure and team performance: the case of English Premier League soccer teams. *Soc. Netw.* **34**(4), 682–690 (2012). <https://doi.org/10.1016/j.socnet.2012.08.004>
2. Ievoli, R., Palazzo, L., Ragozini, G.: On the use of passing network indicators to predict football outcomes. *Knowl.-Based Syst.* **222**, 106997 (2021). <https://doi.org/10.1016/j.knosys.2021.106997>
3. Ievoli, R., Gardini, A., Palazzo, L.: The role of passing network indicators in modeling football outcomes: an application using Bayesian hierarchical models. *AStA Adv. Stat. Anal.* **107**(1–2), 153–175 (2023). <https://doi.org/10.1007/s10182-021-00411-x>
4. Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: an innovative method to cluster networks. *Stat. Model.* **19**(1), 28–54 (2019). <https://doi.org/10.1177/1471082X18808628>
5. Palazzo, L., Ievoli, R., Ragozini, G.: Testing styles of play using triad census distribution: an application to men’s football. *J. Quant. Anal. Sports* **19**(2), 125–151 (2023). <https://doi.org/10.1515/jqas-2022-0010>
6. Håland, E.M., Wiig, A.S., Hvattum, L.M., Stålhane, M.: Evaluating the effectiveness of different network flow motifs in association football. *J. Quant. Anal. Sports* **16**(4), 311–323 (2020). <https://doi.org/10.1515/jqas-2019-0097>
7. Faust, K.: Very local structure in social networks. *Sociol. Methodol.* **37**(1), 209–256 (2007). <https://doi.org/10.1111/j.1467-9531.2007.00179.x>
8. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge (1994)



Transition Towards Environmental Sustainability: Evaluating the Role of Green Investment, Environmental Policy, and Financial Development on the Load Capacity Factor in G-7 Countries

Aamir Javed  and Agnese Rapposelli  

University “G. D’Annunzio” of Chieti-Pescara, Viale Pindaro 42, Pescara, Italy
{aamir.javed, agnese.rapposelli}@unich.it

Abstract. This empirical study considers the load capacity factor, a more comprehensive indicator that takes into account both the demand and supply sides to measure environmental quality, and explores the role of green investment, environmental policy stringency, financial development, natural resources rent and trade openness in the context of the load capacity curve hypothesis for G-7 countries. To this purpose, this work employs the cross-sectional autoregressive distributed lag (CS-ARDL) model to investigate the impact of each variable. The results obtained suggest that green investment and environmental policy stringency both contribute to improve environmental quality by expanding the load capacity factor, while financial development and natural resource rent significantly reduce the load capacity factor and enhance environmental degradation. Additionally, the findings also validate the load capacity curve hypothesis in the G-7 context. Hence, the present study highlights that green investment and policy stringency are necessary to achieve ecological sustainability. Besides, it presents in-depth ecological solutions so that G-7 economies can implement finalized policy actions to reach the targets set in COP26.

Keywords: Green investment · Environmental policy stringency · Load capacity factor · CS-ARDL model

1 Introduction

Ecological degradation and climate change have become a constant threat to sustainable development. The economic growth occurred in past decades is backed by rapid industrialization, consumption of fossil fuels, and unconscious use of natural resources, resulting both in a major climate shift and resources inequality that are endangering the existence of humankind on earth. Hence, a significant policy shift is required in the driver of economic growth to save the planet for coming generations [8, 14]. Considering this, world leaders are making many legislative and academic efforts to curtail the exacerbation of ecological degradation. There exist many international frameworks

and agreements pertaining to climate change at a global level, including the Sustainable Development Goals (SDGs), the Paris Agreement, and the 26th United Nations Climate Change Conference (COP26). The objectives set in the COP26 summit include achieving global “Net Zero” emissions by the year 2050, ensuring that the increase in the global average temperature should remain below 1.5 degrees Celsius, safeguarding natural habitats and communities, mobilizing sustainable finance, and fostering international cooperation to effectively execute these objectives [6, 7]. Nevertheless, the latest reports on SDGs for the year 2022 reveal that several economically developed countries, even those belonging to the G-7 economies, are failing to meet the objectives set [16]. Here the question arises about the effectiveness of policy choices and actions relative to green research and development, environmental regulations, and financial efficacy in controlling environmental damages and pushing sustainability limits.

Based on the above discussion, in our opinion it is important to investigate the impact of green investment, environmental policy stringency, financial development, natural resources, and trade activities on the load capacity factor (LCF). Hence, the main question this work ought to answer is “*whether green investment, environmental policy stringency, natural resources, and financial development really affect the index of load capacity factor for G-7 economies?*”. This empirical study aims to answer the above question in the framework of the load capacity curve (LCC) hypothesis by analyzing the extent to which financial, green investment, environmental policy stringency, financial development, natural resources, and trade activities contribute to increase the load capacity factor of G-7 countries. To this purpose, we apply the cross-sectional autoregressive distributed lag (CS-ARDL) model on panel data covering the period 1990–2019. After investigating the previous literature on the topic, we identified several necessary justifications for the novel objectives of our study. Firstly, this research considers the LCF as a measure of climate change and resilience under the theoretical framework of the LLC hypothesis. Most of the previous studies have focused on carbon emissions and ecological footprint [3, 9, 12, 14], which only consider one aspect (demand side factors) of ecological sustainability. The other important aspect of natural resource regeneration and bio-capacity (supply-side factors) of the earth has been ignored. Secondly, this study includes the most important determinants to understand their possible role in the LCF for G-7 countries.

This paper is organized as follows. Section 2 provides the relevant literature, Sect. 3 presents the data and the methodology used, Sect. 4 lists and discusses the results obtained and Sect. 5 concludes and provides some policy implications.

2 Literature Review

The need for green financing and investments is being highlighted in every forum to mobilize the involved parties, as green investment represents the key to green innovation, clean energy, and green growth, which ultimately results in securing the biocapacity of Earth. In this context, [13] examined the impact of green investment on China’s environmental quality. Similarly, [11] investigated the scenario of green investments linked to CO₂ emissions in European countries. Their results revealed that green investment tends to improve environmental quality. By focusing on the role of R&D, [5] analyzed its influence on the LCF of G-7 economies. Their findings showed the existence of the LLC

hypothesis. Exploring the role of some important indicators, [1] investigated the impact of green taxation and green innovation for emerging economies. Their results revealed the U-shape relationship between ecological sustainability and economic growth. Focusing on renewable and nuclear energy R&D investment, [10] investigated the LLC hypothesis for Japan, showing that renewable energy R&D expenditures promotes ecological sustainability. Considering the financial perspective, [2] studied the effect of financial globalization on the LCF for India, revealing its positive impact. Besides, according to [3], investments in green technological innovation represents a suitable policy choice to curb consumption-based air pollution in G-7 economies. Discussing the role of environmentally related technologies in economic development and renewable energy utilization, [15] stated that investment in environmentally related technologies improves the environmental quality of APEC countries. However, there are some studies that did not prove the existence of the LLC hypothesis and showed that income growth is not the cure for low LCF. For instance, [12] investigated this problem by including renewable and nuclear energy R&D expenditures in the context of Germany. By applying different econometric techniques, the authors reported the validity of EKC hypothesis instead of LCC hypothesis.

3 Data and Method

In order to achieve the objective of the study, annual panel time series data is utilized for the G-7 countries. Table 1 lists the variables considered along with their symbols, descriptions, and data sources.

Table 1. Description of variables

Variables	Symbol	Unit of measurement	Sources
Load capacity factor	LCF	Biocapacity divided by the ecological footprint (in global hectares per person)	GFN
Green investment index (PCA)	GI	i) Environmentally related governmental R&D budget (% of total governmental R&D budget); ii) Renewable energy R&D budget (% of total energy R&D budget); iii) Development of environmentally friendly technologies (% of all technologies)	OECD
Financial development	FD	Financial development index	IMF
Environmental policy	EP	Environmental policy stringency index	OECD
Economic growth	GDP	GDP per capita (constant 2010 US\$)	WDI
Natural resources	NRR	Natural resources rent (% of GDP)	WDI
Trade openness	TRD	Sum of exports and imports on GDP	WDI

To analyze the relationship between the variables mentioned above and the load capacity factor the following model is designed:

$$LCF_{it} = \beta_0 + \beta_1 GI_{it} + \beta_2 EPS_{it} + \beta_3 FD_{it} + \beta_4 NRR_{it} + \beta_5 TR_{it} + \beta_6 GDP_{it} + \beta_7 GDP_{it}^2 + \varepsilon_{it} \quad (1)$$

To estimate the long-run dynamics among the variables, this study utilizes the cross-sectional autoregressive distributed lag (CS-ARDL) model developed by [4]. The formulation of the model is the following one:

$$LCF_{i,t} = \sum_{i=0}^{p_u} \alpha_{i,t} LCF_{i,t-1} + \sum_{i=0}^{p_v} \delta_{i,t} V_{i,t} + \varepsilon_{it} \quad (2)$$

$$LCF_{i,t} = \sum_{i=0}^{p_u} \alpha_{i,t} LCF_{i,t-1} + \sum_{i=0}^{p_v} \delta_{i,t} \bar{V}_{i,t} + \sum_{i=0}^{p_w} \gamma_i \bar{W}_{t-1} + \varepsilon_{i,t} \quad (3)$$

where:

$$\bar{W}_{t-1} = (\bar{LCF}_{t-1}, \bar{V}_{t-1}).$$

$\bar{W}_{t-1}$ represents the average of $\bar{LCF}_{t-1}$ (dependent variable) and $\bar{V}_{t-1}$ (all the explanatory variables) which overcomes the cross-sectional dependence problem; p_u , p_v , p_w represent the lag values. In the CS-ARDL estimation approach, the estimates of the short-run estimates are used to compute the long-term estimates.

4 Results and Discussion

The estimated results of the CS-ARDL model applied show that the coefficient associated with green investment is positive and statistically significant at the level of 10%, indicating that green investment has an encouraging impact on the LCF. To be more precise, a 0.015% increase in the LCF level is attributed to a variation equal to 1% in green investment in the long-term. This result illustrates that green investment plays a constructive role in the ecological sustainability of G-7 countries during the period analyzed. Similarly, the results obtained indicate that environmental policy stringency in G-7 countries is positive and significantly linked with the LCF level (a 1% increase in environmental policy stringency causes an increase in LCF equal to 0.023% in the long run), thus highlighting that the implementation of environmental policy in G-7 countries can improve environmental health. On the contrary, the findings of this work indicate that natural resources rent has an adverse impact on the environmental quality in G-7 nations. More precisely, the results obtained demonstrate that a 1% increase in natural resources rent tends to decrease the LCF level by 0.023%. Furthermore, the estimated coefficients also show that both financial development and trade openness exert a negative influence on the LCF (a 1% increase in financial development and trade openness cause a decrease in the LCF level equal to 0.042% and 0.044%, respectively) (Table 2).

Table 2. CS-ARDL results.

	Long-run estimates				Short-run estimates			
	Coeff	Std. E	Z	P > Z	Coeff	Std. E	Z	P > Z
GI	0.015*	0.008	1.70	0.090	0.010*	0.005	1.82	0.069
EPS	0.023*	0.013	1.81	0.070	0.022*	0.012	1.72	0.085
FD	−0.042***	0.105	−0.40	0.003	−0.022**	0.085	−0.26	0.054
NRR	−0.023**	0.011	−2.13	0.033	−0.019**	0.008	−2.20	0.028
GDP	−18.454**	8.044	−2.29	0.022	−14.247**	6.719	−2.12	0.034
GDP2	0.866**	0.372	2.32	0.020	0.667**	0.310	2.15	0.031
TRD	−0.044***	0.074	−0.60	0.005	−0.045***	0.063	−0.71	0.000
ECM	-	-	-	-	−0.861***	0.084	-	0.000

Note: ***, ** and * denote significance level at 1%, 5% and 10%

5 Conclusion

The present study explores the role of green investment and environmental policy stringency on the LCF in the seven most developed nations in the world. Additionally, the influence of financial sector development, natural resources rent, and trade openness in the framework of LCC theory is also investigated, with a specific focus on the thirty years spanning from 1990 to 2019. The dynamic connections between the research variables are estimated using the CS-ARDL method. The empirical results show that green investment and environmental policy stringency positively and significantly influence the LCF both in the long- and short-run. On the contrary, financial development, natural resources rent, and trade activities have a significant negative relationship with the LCF and represent the main sources of worsening environmental quality in the countries analyzed. The findings of our research are crucial and have insightful policy recommendations for the governments of G-7 countries that will help them achieve the SDGs and meet the objectives set in the COP26 context.

Funding. This work was funded by EU – NextGenerationEU – Call Prin 2022 PNRR, D.D. 1409 14.9.2022, National Recovery and Resilience Plan (PNRR) Mission 4 – Component 2, Investment n. 1.1 (project code P2022XME5P titled “Circular Economy from the Mathematics for Signal Processing prospective”, grant number E53D23018040001).

References

1. Afshan, S., Yaqoob, T.: Unravelling the efficacy of green innovation and taxation in promoting environmental quality: a dual-model assessment of testing the LCC theory in emerging economies. *J. Cleaner Product.* **416**, 137850 (2023). <https://doi.org/10.1016/j.jclepro.2023.137850>
2. Akadiri, S.S., Adebayo, T.S., Riti, J.S., Awosusi, A.A., Inusa, E.M.: The effect of financial globalization and natural resource rent on load capacity factor in India: an analysis using the dual adjustment approach. *Environ. Sci. Pollut. Res.* **29**(59), 89045–89062 (2022)

3. Bai, J., Han, Z., Rizvi, S.K.A., Naqvi, B.: Green trade or green technology? The way forward for G-7 economies to achieve COP 26 targets while making competing policy choices. *Technol. Forecast. Soc. Chang.* **191**, 122477 (2023)
4. Chudik, A., Pesaran, M.H.: Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *J. Econometrics* **188**(2), 393–420 (2022)
5. Dogan, A., Pata, U.K.: The role of ICT, R&D spending and renewable energy consumption on environmental quality: testing the LCC hypothesis for G7 countries. *J. Clean. Prod.* **380**, 135038 (2022)
6. IEA: Net Zero by 2050. A Roadmap for the Global Energy Sector. IEA Publications, France (2021a)
7. IEA: World Energy Outlook 2021. IEA Publications, France (2021b)
8. Jiakui, C., Abbas, J., Najam, H., Liu, J., Abbas, J.: Green technological innovation, green finance, and financial development and their role in green total factor productivity: empirical insights from China. *J. Cleaner Product.* **382**, 135131 (2023)
9. Jiang, Y., Hossain, M.R., Khan, Z., Chen, J., Badeeb, R.A.: Revisiting research and development expenditures and trade adjusted emissions: green innovation and renewable energy R&D role for developed countries. *J. Knowl. Econ.* **15**, 2156–2191 (2023)
10. Kartal, M.T., Pata, U.K., Destek, M.A., Caglar, A.E.: Environmental effect of clean energy research and development investments: evidence from Japan by using load capacity factor. *J. Clean. Prod.* **416**, 137972 (2023)
11. Lyeonov, S., Pimonenko, T., Bilan, Y., Štreimikiene, D., Mentel, G.: Assessment of green investments' impact on sustainable development: Linking gross domestic product per capita, greenhouse gas emissions and renewable energy. *Energies* **12**(20), 3891 (2019)
12. Pata, U.K., Kartal, M.T., Erdogan, S., Sarkodie, S.A.: The role of renewable and nuclear energy R&D expenditures and income on environmental quality in Germany: scrutinizing the EKC and LCC hypotheses with smooth structural changes. *Appl. Energy* **342**, 121138 (2023)
13. Shen, Y., Su, Z.W., Malik, M.Y., Umar, M., Khan, Z., Khan, M.: Does green investment, financial development and natural resources rent limit carbon emissions? A provincial panel analysis of China. *Sci. Total Environ.* **755**, 142538 (2021)
14. Shu, H., Wang, Y., Umar, M., Zhong, Y.: Dynamics of renewable energy research, investment in EnvoTech and environmental quality in the context of G7 countries. *Energy Economics* **120**, 106582 (2023)
15. Sun, Y., Guan, W., Razzaq, A., Shahzad, M., An, N.B.: Transition towards ecological sustainability through fiscal decentralization, renewable energy and green investment in OECD countries. *Renew. Energy* **190**, 385–395 (2022)
16. World Bank: Tracking SDG7: The Energy Progress Report. World Bank, Washington DC (2022)



Offshore Wind Farm Development in Italy

Feroz Khan  and Agnese Rapposelli  

University “G. D’Annunzio” of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy
{feroz.khan, agnese.rapposelli}@unich.it

Abstract. High prices of fossil fuels have resulted in high prices of electricity in recent times. This phenomenon have made renewables, such as wind power, a more feasible option. In particular, offshore wind turbines are considered a more suitable alternative to onshore ones due to socio-economic, environmental, and technical benefits. However, the deployment of wind farms requires appropriate planning for the project to be feasible. In this context, Italy is looking forward to enhance its offshore wind potential. However, strict permitting procedures and guidelines for these projects is rather a challenge for the country. The Strait of Sicily is considered to be a suitable region for the deployment of such wind farms. Hence, this study focuses on the problem of sustainable sites selection for floating offshore wind development in the case of Sicily. To this purpose, this work employs the fuzzy based Multi-Criteria Decision Making (MCDM) approach together with the Geographic Information Systems (GIS) software to identify optimal locations for future wind farm development in the Strait of Sicily.

Keywords: Offshore wind farms · Sicily · fuzzy Multi-Criteria Decision Making (MCDM) approach · Geographic Information Systems (GIS)

1 Introduction

Energy security has become an important topic, especially after the Russian-Ukraine conflict. European countries have been the most affected by this crisis because they rely on the import of natural gas from Russia. Hence, such countries are driven to diversify their energy mix to be more energy autonomous. This is also in line with Sustainable Development Goal 7 (SDG 7 – “*Affordable and clean energy*”) which aims at achieving “*affordable, reliable, sustainable and modern energy for all*”. Besides, the “REPowerEU” plan, one of the European Union strategies to reduce dependence on Russian fossil fuels and fast forward the green transition, focuses on the same goal [1]. Amongst the renewable energy sources, wind energy has received the most attention in recent times. Italy, which represents the third largest economy of the EU is fostering the growth of renewable energy sources to be more energy autonomous. Recently, the offshore wind industry has become popular because of the increased interest of wind power developers in floating-based foundations. Terna, which is the Transmission System Operator (TSO) in Italy, has received several requests from wind developers to locate their future projects along the coasts of Sardinia and Sicily. In particular, Sicily is receiving more interest

from them, as evidenced by projects like “Med Wind” which has an installed capacity of about 2.8 GW. However, the most challenging phase in the life cycle of these projects is the identification of appropriate zones for wind farm deployment. These projects must undergo several environmental, socio-economic and technical obstacles to be able to go through the permitting phases, especially in the case of Italy where these procedures have a track record of being delayed. Since no suitability assessment for these sites has been explored in the literature for the Italian context, the current study fills in this gap by focusing on sustainable sites selection for offshore wind farm development along the west coast of Sicily. To this purpose, this work employs a rather novel Multi-Criteria Decision Making (MCDM) approach known as Full Consistency Method (FUCOM) to allocate weights to the spatial criteria. The fuzzy set theory is also employed to address the notion of ambiguity during the decision-making process. Finally, the Geographic Information Systems (GIS) software is used to perform the analysis and eventually help determine the suitability map for Sicily.

The study unfolds as follows. Section 2 introduces the literature review, Sect. 3 presents the data and the methodology used, Sects. 4 lists the results and Sect. 5 concludes the paper.

2 Literature Review

Site selection is the most challenging step during the planning phase of offshore wind farm development. The three most important financial parameters to determine the feasibility of this type of wind farms include distance to shore and ports and bathymetry. Life cycle costs models are one of the approaches able to help in such decision-making problems [2]. The feasibility determination of offshore wind farms has been studied based on evaluating scenarios for the years 2030 and 2060 for Italy. The evaluation is performed based on the criterion of distance from shore, and the economic models used are based on the Levelized Cost of Energy (LCOE). Results suggest that there is a considerable potential of offshore wind power in Italy; however, there are several challenges because the technology is not yet mature and planning phases require adequate multi-disciplinary techniques [3]. According to recent literature, GIS-based multi criteria decision-making approach is the most rational way of dealing with such problems; more specifically, the Weighted Overlay technique is applied in GIS for these site selection problems [4]. In addition, to evaluate the decision criteria, most literature employs the Analytical Hierarchy Process (AHP) technique. The decision criteria in these cases are spatial data which are cleaned in the GIS software to perform the weighted overlay analysis to determine optimal sites based on different levels of suitability [5]. [6] have carried out a site selection for floating offshore wind farm development in the Aegean Sea (Greece). Nine evaluation criteria have been considered in this case for determining the suitability map; the criteria are divided in terms of exclusion and inclusion factors. Inclusion criteria are considered to be important in determining the optimal location whereas exclusion criteria are considered to be detrimental to the project. The selection of the type of criteria is often categorized under social, economic, environmental and technical aspects. Comprehensive consideration of these criteria is evident in the case of Morocco which employs the AHP technique to allocate weights to such spatial criteria [7].

3 Data Collection and Methodology

The GIS and MCDM tools are some of the most employed methodologies in the literature to perform a site suitability analysis, as mentioned in the previous section. The hybrid approach helps accommodating both secondary and primary datasets to tackle such problems. Secondary data in these cases correspond to the gathering of spatial data based on environmental, socio-economic and technical aspects, whilst primary data are handled through the application of the MCDM approach where the evaluation criteria are assessed with the help of experts. Finally, the weighted overlay tool in ArcGIS software is employed to perform site suitability assessment. In this study, the spatial data layers for 15 evaluation and exclusion criteria are gathered through an extensive literature review. The evaluation criteria determine the feasibility of the project, while the exclusion criteria are represented by risk factors that are considered as buffer zones. To conform to the Italian regulations, this type of data is primarily based on the guidelines provided by the National Wind Energy Association (ANEV). Spatial data layers collected are in raster and vector format which is prepped in the ArcGIS software for further analysis. To perform raster analysis, it is important to convert vector data sets into their raster format. The fuzzy FUCOM approach allocates weights to the evaluation criteria which are subsequently multiplied by the corresponding raster layers. Finally, all the raster layers are added together using the weighted overlay tool to yield the most suitable zones.

The suitability index (SI) and the associated weighted linear sum are explained in the following equation:

$$SI = \sum_{j=1}^n W_j \times x_{ij} \quad (1)$$

where W_j denotes the importance weight of criteria j and x_{ij} denotes the standardized score of the raster layers corresponding to cell i and criteria j .

The methodological framework of the study is shown in Fig. 1 below.

4 Results

The results of this study can be presented in two steps. The first step involves the results derived from the fuzzy FUCOM approach. The experts allocated the most weights to wind speed and distance from shore both of which contribute towards the economic and technical feasibility of the project. The ArcGIS software is employed hereafter to perform the weighted overlay analysis to yield the final suitability map. The suitability map is composed of five classes ranging from the least suitable zones to the most suitable zones. The most suitable sites can be considered by both the government and wind developers to locate their floating offshore wind turbines along the Strait of Sicily. The suitability map is shown in Fig. 2 below.

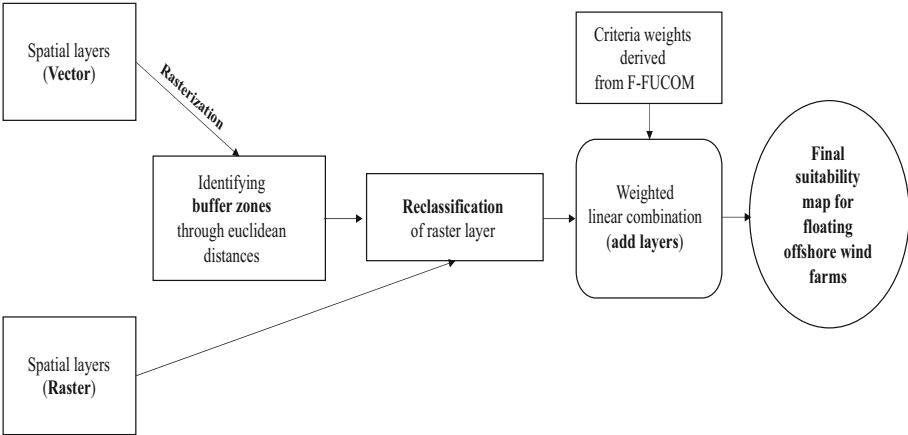


Fig. 1. Methodological framework of the work.

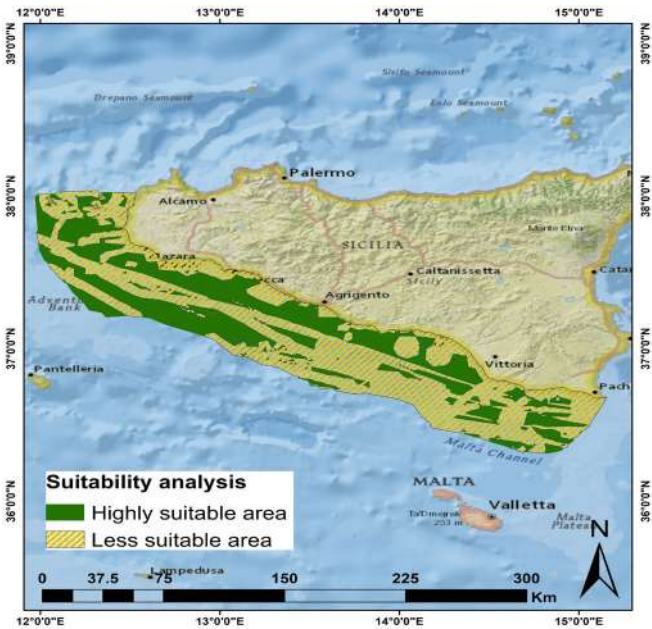


Fig. 2. Suitability map

5 Conclusion

The concept of floating offshore wind turbines is still novel in Italy. The only offshore wind farm in the country, based on a fixed foundation platform, is located in the Strait of Sicily. Hence, no sites suitability assessment for floating foundation-based offshore wind farms has been carried out in the literature for the Italian context. To this purpose,

this study has involved evaluation and exclusion criteria that are specifically based on the dynamics of Italy. The application of the fuzzy FUCOM approach, which is considered to be more reliable as compared to the traditional AHP and Best Worst Method (BWM), further strengthened the results of the final suitability map. In addition, the fuzzy set theory applied to the decision-making process enhanced the validity of the data collected from experts in the energy field.

The employed methodology and the secondary sources used to obtain the evaluation and exclusion criteria can be used for future studies to perform a similar analysis for other Italian regions, such as Sardinia and Apulia.

References

1. European Commission: 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 – REPowerEU Plan. European Commission, Brussels (2022).
2. Maienza, C., Avossa, A.M., Picozzi, V., Ricciardelli, R.: Feasibility analysis for floating offshore wind energy. *Int. J. Life Cycle Assess.* **27**(1), 796–812 (2022)
3. Serri, L., Colle, L., Vitali, B., Bonomi, T.: Floating offshore wind farms in Italy beyond 2030 and beyond 2060: preliminary results of a techno-economic assessment. *Appl. Sci.* **10**(24), 8899 (2020)
4. Zhou, X., et al.: Site selection for hybrid offshore wind and wave power plants using a four-stage framework: a case study in Hainan, China. *Ocean Coastal Manage.* **218**(1), 106035 (2022)
5. Gkeka-Serpetsidaki, P., Tsoutsos, T.: A methodological framework for optimal siting of offshore wind farms: a case study on the island of Crete. *Energy* **239D**, 122296 (2022)
6. Stefanakou, A.A., Nikitakos, N., Lilas, T., Pavlogeorgatos, G.: A GIS-based decision support model for offshore floating wind turbine installation. *Int. J. Sustain. Energ.* **38**(7), 673–691 (2019)
7. Taoufik, M., Fekri, A.: GIS-based multi-criteria analysis of offshore wind farm development in Morocco. *Energ. Convers. Manage.* **11**(1), 100103 (2021)



Joint Complete Records Occurrence over Farlie-Gumbel-Morgenstern Sequences

Amir Khorrami Chokami^{1(✉)} and Silvia Montagna²

¹ Dipartimento di Matematica e Informatica, Università di Cagliari,
Via dell'Ospedale 72, 09124 Cagliari, Italia
amir.khorramichokami@unica.it

² Dipartimento di Economia “Marco Biagi”, Università degli Studi di Modena e
Reggio Emilia, via Berengario 51, 41121 Modena, Italy

Abstract. A record over a sequence of iid absolutely continuous random variables is defined as a random variable surpassing all preceding ones. The extension to the multivariate case is not obvious. In this project, we focus on the so-called complete records. We examine the probability of appearance of two complete records at two different times, whenever the within sequences of random vectors where the dependence among components follows the Farlie-Gumbel-Morgenstern copula.

Keywords: Complete Records · Time Dependence · Standard Max Stable distribution · FGM copula

1 Introduction

In recent times, there has been a growing interest on the occurrence of extreme events. The study of maxima in a multivariate context presents the challenge of tackling the dependence among extreme events, in connection with the parent copula of the underlying sequence of random vectors (see [9, 10] among others). Let us consider now a sequence of iid random variables $X_1, X_2, \dots$. The random variable X_m is called a *record* if $X_m > \max(X_1, \dots, X_{m-1})$, $m \geq 2$ (obviously, X_1 is the first record because it is the first variable appearing in the sequence). In this simple case, various results concerning the time of appearance of a record, the asymptotic distribution of the m -th record as $m \rightarrow \infty$, increments between records, are available (see [1], [5, Sections 6.2 and 6.3], [2, 4], among others). In particular, if we call $I_m^R = \mathbf{1}(X_m > \max(X_1, \dots, X_{m-1}))$, $m \in \mathbb{N}$, then

$$\mathbb{P}(I_j^R = 1, I_k^R = 1) = \mathbb{P}(I_j^R = 1) \mathbb{P}(I_k^R = 1) = j^{-1}k^{-1}, \quad j < k, \quad (1)$$

indicating independence between the appearance of records at two different times.

The extension to the multivariate case is not an easy task. The first issue we have to face regards the non-unique definition of a multivariate record. In

this work, we consider the so-called complete records, defined as follows. Let $\mathbf{X}_1, \mathbf{X}_2, \dots$, be iid random vectors in $\mathbb{R}^d$. For any two integers $j < k$, we define

$$\mathbf{M}_j := \max_{1 \leq i \leq j} \mathbf{X}_i, \quad \mathbf{M}_j^k := \max_{j \leq i \leq k} \mathbf{X}_i,$$

where the maximum is taken componentwise. A random vector $\mathbf{X}_m$ is a *complete record* (CR) if each component is a univariate record itself, i.e., $\mathbf{X}_m > \mathbf{M}_{m-1}$. The event indicating the appearance of a complete record at time m is then $I_m^{\text{CR}} := \mathbb{1}(\mathbf{X}_m > \mathbf{M}_{m-1})$. Some findings regarding complete records across sequences of random vectors with independent components and with Marshall-Olkin copula are available in [2, 3, 8], respectively.

This paper studies the left-hand side of Eq. (1) in a bivariate context, precisely when the copula of the underlying sequence of random vectors is of the Farlie-Gumbel-Morgenstern (FGM) type:

$$F(x, y) = e^{(x+y)} (1 + \lambda (1 - e^x) (1 - e^y)), \quad x, y \leq 0 \text{ and } |\lambda| \leq 1, \quad (2)$$

with bivariate density given by

$$f(x, y) = e^{x+y} (1 + \lambda (2e^x - 1) (2e^y - 1)). \quad (3)$$

This copula is valuable due to its simple structure and straightforward interpretation of the parameter λ , which captures dependence characteristics. These qualities have made the FGM distribution a popular choice for various applications, including capital allocation and the so-called collective risk models. For a thorough understanding of this copula in an extreme-value context, refer to [6]. We highlight our findings are independent of the marginal distribution function F_{X_i} , $i = 1, 2$, provided that they are continuous. In the following, we consider a sequence $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots \in \mathbb{R}^2$ of *standard max-stable* random vectors (i.e. with Negative-Exp(1) margins and such that $\mathbf{M}_n \stackrel{d}{=} \boldsymbol{\eta}_1/n$) and FGM copula. This is also the framework of [7], which deals with single complete records.

2 Complete Records Occurrences

Theorem 1. *Let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ be a sequence of bivariate standard max-stable random vector with FGM copula. Then, the probability of appearance of two complete records at times j and k , with $j < k$, is*

$$\mathbb{P}(I_j^{\text{CR}} = 1, I_k^{\text{CR}} = 1) = \sum_{i=1}^{16} p_i \quad (4)$$

for $j, k \in \mathbb{N}, j < k$. The p_i 's are given in the proof for readability.

Proof. For what concerns the probability of the joint events we have

$$\begin{aligned}
& \mathbf{P}(I_j^{\text{CR}} = 1, I_k^{\text{CR}} = 1) = \mathbf{P}(\boldsymbol{\eta}_j > \mathbf{M}_{j-1}, \boldsymbol{\eta}_k > \mathbf{M}_{k-1}) \\
& = \mathbf{P}\left(\boldsymbol{\eta}_j > \frac{\boldsymbol{\eta}_1}{j-1}, \boldsymbol{\eta}_k > \max\left(\boldsymbol{\eta}_j, \frac{\boldsymbol{\eta}_2}{k-j-1}\right)\right) \\
& = \int_{(-\infty, 0]^2} \int_{(-\infty, \mathbf{y}]} \mathbf{P}(\boldsymbol{\eta}_1 < (j-1)\mathbf{x}, \boldsymbol{\eta}_2 < (k-j-1)\mathbf{y}) \, d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{x}) \, d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{y}) \\
& = \int_{(-\infty, 0]^2} \mathbf{P}(\boldsymbol{\eta}_2 < (k-j-1)\mathbf{y}) \int_{(-\infty, \mathbf{y}]} \mathbf{P}(\boldsymbol{\eta}_1 < (j-1)\mathbf{x}) \, d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{x}) \, d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{y}).
\end{aligned}$$

Let us compute the internal integral, which will yield a function of $\mathbf{y}$.

$$\begin{aligned}
& \int_{(-\infty, \mathbf{y}]} \mathbf{P}(\boldsymbol{\eta}_1 < (j-1)\mathbf{x}) \, d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{x}) \\
& = \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} \mathbf{P}\left(\eta_1^{(1)} < (j-1)x_1, \eta_1^{(2)} < (j-1)x_2\right) \, d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{x}) = \\
& = \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} e^{(j-1)(x_1+x_2)} \left(1 + \lambda \left(1 - e^{(j-1)x_1}\right) \left(1 - e^{(j-1)x_2}\right)\right) \cdot \\
& \quad \cdot e^{x_1+x_2} \left(1 + \lambda (2e^{x_1} - 1) (2e^{x_2} - 1)\right) \, dx_1 \, dx_2 = \\
& = \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} e^{j(x_1+x_2)} \, dx_1 \, dx_2 \\
& \quad + \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} e^{j(x_1+x_2)} \lambda \left(1 - e^{(j-1)x_1}\right) \left(1 - e^{(j-1)x_2}\right) \, dx_1 \, dx_2 \\
& \quad + \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} e^{j(x_1+x_2)} \lambda (2e^{x_1} - 1) (2e^{x_2} - 1) \, dx_1 \, dx_2 \\
& \quad + \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} e^{j(x_1+x_2)} \lambda^2 \left(1 - e^{(j-1)x_1}\right) \left(1 - e^{(j-1)x_2}\right) (2e^{x_1} - 1) (2e^{x_2} - 1) \, dx_1 \, dx_2 \\
& =: I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

It can be shown that

$$\begin{aligned}
I_1 &= \frac{1}{j^2} e^{j(y_1+y_2)} \\
I_2 &= \lambda \prod_{i=1}^2 \left(\frac{e^{jy_i}}{j} - \frac{e^{(2j-1)y_i}}{2j-1} \right) \\
I_3 &= \lambda \prod_{i=1}^2 \left(2 \frac{e^{(j+1)y_i}}{j+1} - \frac{e^{jy_i}}{j} \right) \\
I_4 &= \lambda^2 \prod_{i=1}^2 \left(\frac{2}{j+1} e^{(j+1)y_i} - \frac{e^{jy_i}}{j} - \frac{e^{2jy_i}}{j} + \frac{e^{(2j-1)y_i}}{2j-1} \right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
 & \int_{(-\infty, 0]^2} \mathbf{P}(\boldsymbol{\eta}_2 < (k-j-1)\mathbf{y}) \int_{(-\infty, \mathbf{y}]} \mathbf{P}(\boldsymbol{\eta}_1 < (j-1)\mathbf{x}) d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{x}) d\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{y}) \\
 &= \int_{-\infty}^0 \int_{-\infty}^0 \left[\sum_{i=1}^4 I_i \right] \cdot F((k-j-1)y_1, (k-j-1)y_2) f(y_1, y_2) dy_1 dy_2 \\
 &== \sum_{i=1}^4 \int_{-\infty}^0 \int_{-\infty}^0 I_i \cdot F((k-j-1)y_1, (k-j-1)y_2) f(y_1, y_2) dy_1 dy_2,
 \end{aligned}$$

where F and f are as in Eq.(2) and Eq.(3), respectively. Solving the bivariate integral above, requires the computation of the following 16 integrals, each one relative to a different p_i in Eq.(4). This is because $F((k-j-1)y_1, (k-j-1)y_2) f(y_1, y_2)$ is a sum of four different contributes, each of them being multiplied with each one of the 4 integrals.

For $F \cdot f \cdot I_1$: The corresponding double integral is the sum of the following

$$\begin{aligned}
 p_1 &= \frac{1}{j^2 k^2} \\
 p_2 &= \frac{\lambda}{j^2} \left(\frac{1}{k} - \frac{1}{2k-j-1} \right)^2 \\
 p_3 &= \frac{\lambda}{j^2} \left(\frac{2}{k+1} - \frac{1}{k} \right)^2 \\
 p_4 &= \frac{\lambda^2}{j^2} \left(\frac{2}{k+1} - \frac{1}{k} - \frac{2}{2k-j} + \frac{1}{2k-j-1} \right)^2
 \end{aligned}$$

For $F \cdot f \cdot I_2$: The corresponding double integral is the sum of the following

$$\begin{aligned}
 p_5 &= \lambda \left(\frac{1}{jk} - \frac{1}{(2j-1)(k+j-1)} \right)^2 \\
 p_6 &= \lambda \left(\frac{1}{jk} - \frac{1}{(2j-1)(k+j-1)} - \frac{1}{j(2k-j-1)} - \frac{1}{(2j-1)(2k-3)} \right)^2 \\
 p_7 &= \lambda \left(\frac{2}{j(k+1)} - \frac{2}{(2j-1)(k+j)} - \frac{1}{jk} - \frac{1}{(2j-1)(k+j-1)} \right)^2 \\
 p_8 &= \lambda^2 \left(\frac{2}{j(k+1)} - \frac{2}{(2j-1)(k+j)} - \frac{1}{jk} + \frac{1}{(2j-1)(k+j-1)} \right. \\
 &\quad \left. - \frac{2}{j(2k-j)} + \frac{2}{(2j-1)(2k-1)} + \frac{1}{j(2k-j-1)} - \frac{2}{(2j-1)(2k-3)} \right)^2
 \end{aligned}$$

For $F \cdot f \cdot I_3$: The corresponding double integral is the sum of the following

$$\begin{aligned}
 p_9 &= \lambda \left(\frac{2}{(j+1)(k+1)} - \frac{1}{jk} \right)^2 \\
 p_{10} &= \lambda \left(\frac{2}{(j+1)(k+1)} - \frac{1}{jk} - \frac{2}{(j+1)(2k-j)} + \frac{1}{j(2k-j-1)} \right)^2 \\
 p_{11} &= \lambda \left(\frac{4}{(j+1)(k+2)} - \frac{2}{j(k+1)} - \frac{2}{(j+1)(k+1)} + \frac{1}{jk} \right)^2 \\
 p_{12} &= \lambda^2 \left(\frac{4}{(j+1)(k+2)} - \frac{2}{j(k+1)} - \frac{2}{(j+1)(k+1)} + \frac{1}{jk} \right. \\
 &\quad \left. - \frac{k}{(j+1)(2k-j+1)} + \frac{2}{j(2k-j)} + \frac{2}{(j+1)(2k-j)} - \frac{1}{j(2k-j-1)} \right)^2
 \end{aligned}$$

For $F \cdot f \cdot I_4$: The corresponding double integral is the sum of the following

$$\begin{aligned}
 p_{13} &= \lambda^2 \left(\frac{2}{(j+1)(k+1)} - \frac{1}{jk} - \frac{1}{j(k+j)} + \frac{1}{(2j-1)(k+j-1)} \right)^2 \\
 p_{14} &= \lambda \left(\frac{2}{(j+1)(k+1)} - \frac{1}{jk} - \frac{1}{j(k+j)} + \frac{1}{(2j-1)(k+j-1)} \right. \\
 &\quad \left. - \frac{2}{(j+1)(2k-j)} + \frac{1}{j(2k-j-1)} + \frac{1}{j(2k-1)} - \frac{1}{(2j-1)(2k-2)} \right)^2 \\
 p_{15} &= \lambda^2 \left(\frac{4}{(j+1)(k+1)} - \frac{2}{j(k+1)} - \frac{2}{j(k+j+1)} + \frac{2}{(2j-1)(k+j)} \right. \\
 &\quad \left. - \frac{2}{(j+1)(k+1)} + \frac{1}{jk} + \frac{1}{j(k+j)} - \frac{1}{(2j-1)(k+j-1)} \right)^2 \\
 p_{16} &= \lambda^2 \left(\frac{4}{(j+1)(k+2)} - \frac{2}{j(k+1)} - \frac{2}{j(k+j+1)} + \frac{2}{(2j-1)(k+j)} \right. \\
 &\quad - \frac{2}{(j+1)(k+1)} + \frac{1}{jk} + \frac{1}{j(k+j)} - \frac{1}{(2j-1)(k+j-1)} \\
 &\quad - \frac{4}{(j+1)(k+1)} + \frac{2}{j(2k-j)} + \frac{1}{jk} - \frac{2}{(2j-1)(2k-1)} \\
 &\quad \left. + \frac{2}{(j+1)(2k-j)} - \frac{1}{j(2k-j-1)} - \frac{1}{j(2k-1)} + \frac{1}{(2j-1)(2k-2)} \right)^2.
 \end{aligned}$$

This concludes the proof.

We remark that the challenge in studying records arises from the immediate complexity of computations required when transitioning from independent cases to those where components are dependent. The FGM copula, albeit with difficulties, enables the derivation of exact formulas for complete records' occurrences.

3 Conclusion

In conclusion, this work addresses the challenge of studying the probability of occurrence of two complete records over independent and identically distributed bivariate sequences of standard max-stable random variables with the FGM copula. The authors suggest the use of mathematical softwares to simplify the expressions given in the proof of Theorem 1. Ongoing projects regard the extension to the case of complete records to sequences characterized by a general copula, and the study of alternative definitions of multivariate records.

References

1. Arnold, B.C., Balakrishnan, N., Nagaraja, H.N.: Records. Wiley Series in Probability and Statistics, Wiley, New York (1998)
2. Falk, M., Khorrami Chokami, A., Padoan, S.: Some results on joint record events. *Stat. Probab. Lett.* **135**, 11–19 (2018). <https://doi.org/10.1016/j.spl.2017.11.011>. <https://www.sciencedirect.com/science/article/pii/S0167715217303656>
3. Falk, M., Khorrami Chokami, A., Padoan, S.A.: On multivariate records from random vectors with independent components. *J. Appl. Probab.* **55**(1), 43–53 (2018). <https://doi.org/10.1017/jpr.2018.4>
4. Falk, M., Khorrami Chokami, A., Padoan, S.A.: Records for time-dependent stationary Gaussian sequences. *J. Appl. Probab.* **57**, 78–96 (2020). <https://doi.org/10.1017/jpr.2019.77>
5. Galambos, J.: The Asymptotic Theory of Extreme Order Statistics, 2nd edn. Krieger, Malabar (1987)
6. Hashorva, E., Hüsler, J.: Extreme values in FGM random sequences. *J. Multivariate Anal.* **68**(2), 212–225 (1999). <https://doi.org/10.1006/jmva.1998.1795>. <https://www.sciencedirect.com/science/article/pii/S0047259X9891795X>
7. Khorrami Chokami, A.: Complete records over independent FGM sequences. *Book of Short Papers CLADAG 2023* (2023)
8. Khorrami Chokami, A., Padoan, S.A.: On multivariate records over sequences of random vectors with Marshall-Olkin dependence of components. IN: *Proceedings of the LI Meeting of the Italian Statistical Society* (2022)
9. Leonetti, P., Khorrami Chokami, A.: The maximum domain of attraction of multivariate extreme value distributions is small. *Electron. Commun. Probab.* **27**, 1–8 (2022). <https://doi.org/10.1214/22-ECP501>
10. Resnick, S.I.: Extreme Values, Regular Variation, and Point Processes. *Applied Probability* vol. 4. Springer, New York (1987). <https://doi.org/10.1007/978-0-387-75953-1>



Deconvolution of the Electrical Load Curve Using Blind Source Separation

Guillaume Koechlin¹(✉), Filippo Bovera², and Piercesare Secchi¹

¹ MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy
{guillaume.koechlin,piercesare.secchi}@polimi.it

² Department of Energy, Politecnico di Milano, Milan, Italy
filippo.bovera@polimi.it

Abstract. The electrical load curve, representing electricity consumption in a power system over time, results from the aggregation of the power demand of the consumers in the system. We propose a method relying on blind source separation and functional data analysis to disaggregate the load curve in consumer groups' contributions. Applied to Italian zones from 2018 to 2022, our method succeeds in identifying clusters of consumers with similar consumption profiles and estimate their contribution to the load.

Keywords: Electrical Load · Blind Source Separation

1 Introduction

The electrical load curve refers to the electrical power consumed by a set of consumption units, as a continuous function of time. At the scale of a region, it is a highly predictable quantity, following a distinct daily, weekly and yearly seasonality. Yet, an exploratory study we conducted on daily load curves of different zones in Italy revealed the existence of a large number of characteristic patterns, showing both spatial and time dependence. Our hypothesis is that this variability in the patterns stems from differing levels of activity of consumer groups sharing a similar and fixed daily load *profile*. Therefore, the diversity in the shapes of the daily load curves that we observe results from differently weighted mixtures of these profiles. We propose a load deconvolution method, relying on blind source separation techniques and functional data analysis, to extract the profiles from these load curves and estimate the contribution to the load of the consumer groups associated to these profiles.

2 Data and Methods

2.1 The Data and Their Functional Representation

The data utilized in this analysis (sourced from Terna S.p.A., the Transmission System Operator (TSO) in Italy [1]) correspond to hourly measurements of the electrical load in Italy, from 2018 to 2022 at the market zone level¹.

¹ North, Centre-North, Centre-South, South, Sicily, Sardinia, Calabria.

During preprocessing, we extracted $n = 11\,653$ daily load vectors each composed of $T = 24$ hourly values. Each vector refers to a day observed in one of the 7 market zones; note however that the 7-th market zone Calabria was added only in 2021.

For any day $i \in \{1, \dots, n\}$, the observed load is the realization of a random vector $L_i \in \mathbb{R}_+^T$. We assume that L_i can be represented as

$$L_{i,j} = X_i(j) + \epsilon_{ij} \quad \forall j \in \{1, \dots, T\} \quad (1)$$

where $X_i : [0, T] \rightarrow \mathbb{R}^+$ is a smooth function; $\epsilon_{i1}, \dots, \epsilon_{iT}$ are assumed to be i.i.d zero-mean noise components.

We estimate $X_1, \dots, X_n$ through a *positive smoothing* procedure using a Fourier basis of 13 basis functions, following, for instance, the least squares method described in Ramsay and Silverman [2]. These n curves are collected in a vector of curves $X = [X_1 \dots X_n] : [0, T] \rightarrow \mathbb{R}_+^n$.

2.2 The Model

We are going to analyze the functional data in X building upon the approach to the *Blind Source Separation* problem developed in [3]. Since our focus is on the shape of the load curves, the total consumption over the day becomes a nuisance. Hence, as a preliminary step we apply an L^1 -normalization to $X_1, \dots, X_n$, ensuring that the integral of each curve over a day is equal to 1.

Each (normalized) curve X_i is assumed to be a mixture of K *profiles* $S_1, \dots, S_K$ - with $K \ll n$ - corresponding to distinct *sources* recognized as *consumer clusters*:

$$\begin{cases} X_1 &= c_{11}S_1 + c_{12}S_2 + \dots + c_{1K}S_K \\ X_2 &= c_{21}S_1 + c_{22}S_2 + \dots + c_{2K}S_K \\ &\vdots \\ X_n &= c_{n1}S_1 + c_{n2}S_2 + \dots + c_{nK}S_K \end{cases} \quad (2)$$

Note that for $i = 1, \dots, n$ and $k = 1, \dots, K$, the coefficient c_{ik} represents the *concentration* of the source S_k in the load function X_i .

The profiles $S_1, \dots, S_K$ and the concentrations c_{ik} 's are constrained to be positive. Additionally, we impose that $S_1, \dots, S_K$ have a unit L^1 -norm, leading to the constraint $c_{i1} + \dots + c_{iK} = 1$, for all $i \in \{1, \dots, n\}$. This condition justifies the term "concentration" for the coefficient c_{ik} . The system of equations (2) can be re-written in a matrix form:

$$X = CS \quad (3)$$

where C is the $n \times K$ matrix of concentrations and $S = [S_1 \dots S_K] : [0, T] \rightarrow \mathbb{R}_+^K$.

The goal is to find a matrix $\hat{C}$ and a vector of functions $\hat{S}$ whose product $\hat{C}\hat{S}$ best approximates the data X . With an L^2 loss (squared error), the problem is

formalized by the following optimization:

$$\begin{aligned}
 (\hat{C}, \hat{S}) = & \arg \min_{\substack{C \in \mathbb{R}_+^{n \times K} \\ S \in \mathcal{C}([0, T], \mathbb{R}_+^K)}} \|X - CS\|_{L^2([0, T], \mathbb{R}_+^n)}^2 \\
 \text{s.t. } & \|S_k\|_{L^1} = 1 \quad \forall k \in \{1, \dots, K\} \\
 & \sum_{k=1}^K c_{ik} = 1 \quad \forall i \in \{1, \dots, n\}.
 \end{aligned} \tag{4}$$

Note that, for a vector of functions $U = [U_1 \dots U_n]': [0, T] \rightarrow \mathbb{R}_+^n$ we define:

$$\|U\|_{L^2([0, T], \mathbb{R}_+^n)} = \left(\sum_{i=1}^n \|U_i\|_{L^2}^2 \right)^{\frac{1}{2}}$$

Once the optimization problem is solved, to “de-normalize” the representation of the curves $X_1, \dots, X_n$, for $i = 1, \dots, n$, we may multiply each coefficient c_{ik} , $k = 1, \dots, K$ with the L^1 norm of X_i computed before the normalisation step.

Relaxing the constraints in (4), the Functional Principal Component Analysis (FPCA) [2] could provide a solution for C and S . However, FPCA has drawbacks: the components (i.e. the profiles S_k ’s) are not necessarily positive, making it challenging to interpret them as meaningful contributions to the target quantity, and the forced orthogonality between the components might not align with the load patterns we aim to discover. FPCA can however be used to (i) estimate the number K of profiles S_k ’s and (ii) provide a reference low-rank matrix approximation for comparison with potential solutions of (4).

In order to solve the problem numerically, we need to find a vector representation for the functions $X_1, \dots, X_n$, ideally preserving “smoothness”, such that we get a matrix representation for X ; we then look for a matrix solution S within this representation. Unfortunately, the positivity constraint does not allow to have a direct vector representation for the original functions (see [2], Chap. 6, on constrained functions). Therefore, we obtain a matrix representation for X by evaluating the functions $X_1, \dots, X_n$ on a common grid of p values spanning the interval $[0, T]$. The dimension $p = 97$ (15 min intervals) was chosen to provide a good balance between computational efficiency and approximation quality. Hence, the optimisation problem (2) becomes:

$$\begin{aligned}
 (\hat{C}, \hat{S}) = & \arg \min_{\substack{C \in \mathbb{R}_+^{n \times K}, S \in \mathbb{R}_+^{K \times p}}} \|X - CS\|_F^2 \\
 \text{s.t. } & \|S_k\|_1 = 1 \quad \forall k \in \{1, \dots, K\} \\
 & \sum_{k=1}^K c_{ik} = 1 \quad \forall i \in \{1, \dots, n\}
 \end{aligned} \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Equation (5) characterizes a constrained variant of the standard *Non-negative Matrix Factorization* (NMF) problem. Among the various iterative algorithms to find a local minimum for the unconstrained NMF, the most popular is the *Multiplicative Updates* (MU) [4]. The initialization obviously being the critical point, our approach involved running the MU algorithm $N = 1000$ times, each time with a different initialization - keeping the same initial value $[\frac{1}{p} \dots \frac{1}{p}]'$ for the rows of S but obtaining each time a different initial value for C by uniformly sampling its rows on the simplex $\mathcal{S}(K)$ - generating a large set of solutions to the unconstrained problem. From this large set of candidate solutions, we selected the solution C^*, S^* that best respects the constraints.

3 Results

A preliminary FPCA (see Fig. 1) suggested to choose $K = 3$, explaining 85% of total variability. The profiles S_1^* , S_2^* , and S_3^* obtained as result of the algorithmic pipeline described in the previous section, are pictured in Fig. 2. The concentrations per zone and type of day² are shown in Fig. 3.

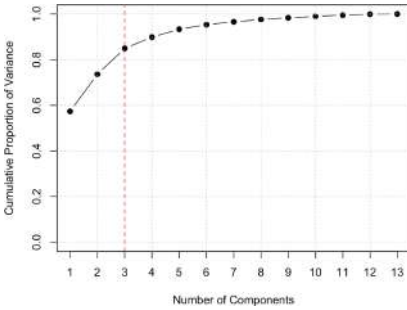


Fig. 1. Scree plot of the FPCA. The elbow point at $K = 3$ PCs is indicated by the vertical dashed red line.

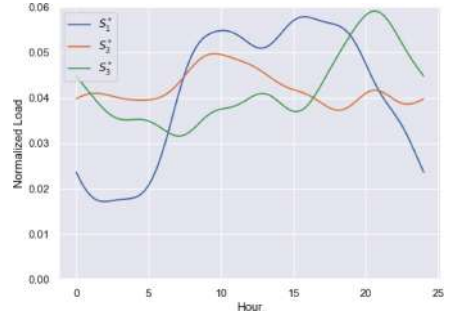


Fig. 2. The profiles S_1^* , S_2^* and S_3^* .

After careful examination, we advance a tentative hypothesis to characterize the consumer clusters associated with the profiles: S_1^* may be associated with the tertiary sector, S_2^* with the industrial sector and S_3^* with the domestic sector.

The estimated contributions of the consumer clusters to each load curve (before normalization) were obtained by de-normalizing the fitted curves in $X^* = C^*S^*$ and retrieving the re-scaled terms $c_{i,1}^*S_1^*$, $c_{i,2}^*S_2^*$ and $c_{i,3}^*S_3^*$. We then computed each cluster's daily consumption per zone by integrating the corresponding contributions. Finally the estimated monthly consumption of each

² Working day (from Monday to Friday), Saturday, and Holiday (Sunday or bank holiday).

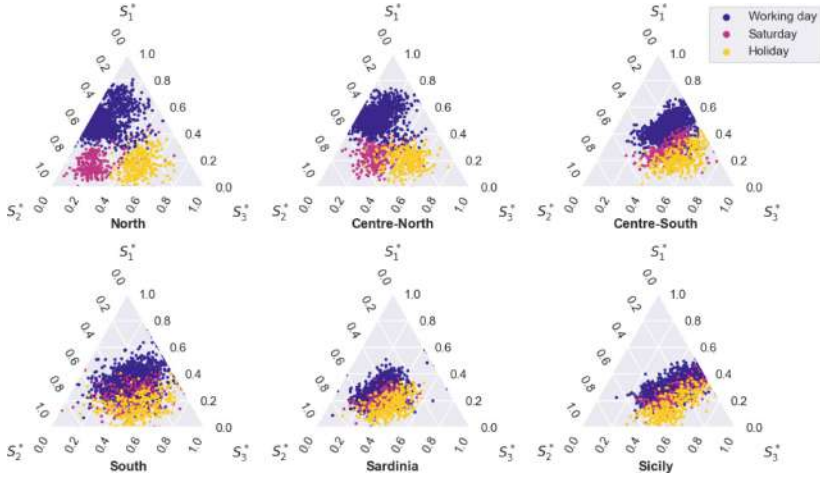


Fig. 3. Scatter plot of the concentrations per zone, grouped by day type, represented on the simplex $\mathcal{S}(3)$. For each day i , the vector of concentrations $[c_{i1} \ c_{i2} \ c_{i3}]'$ belongs to the simplex $\mathcal{S}(3)$. Each vertex of the simplex is associated with one of the three profiles S_1^* , S_2^* , S_3^* ; points closer to a vertex have a higher concentration associated to the corresponding profile. Calabria is not represented as the observations were extremely noisy, possibly due to errors in the data acquisition.

cluster is obtained by summing the daily consumptions. The estimates that correspond to S_1^* along with S_2^* were compared to indicators that Terna use as proxies for the monthly evolution of industrial [6] and tertiary [7] sectors consumption, and appeared to be consistent. We do not have access to ground truth proxies for domestic consumptions (S_3^*). However, the significant increase of consumption for this cluster (Fig. 4) during the COVID-19 lock-down period supports our hypothesis.

4 Conclusion and Discussion

We propose a method to disaggregate the electricity load curve, based on blind source separation techniques. We apply this method to the daily load curves at the market zone level in Italy from 2018 to 2022. The results show that our method is able to individuate profiles corresponding to clusters of consumers characterized by similar load profiles throughout the day, and to identify them. Finally, we show that the proposed method can be used to produce valid estimates of the consumption of these consumer clusters. For the sake of discussion, we here mention some limitations of the method. First, we lack a ground truth validation of the results. This is due to the absence of high frequency load data for the sectors that we believe are corresponding to the profiles S_1^* , S_2^* and S_3^* . Second, the uncertainty of our estimates is not yet addressed. Lastly, our assumption of time and space constant profiles may be too simplistic. Nevertheless, we

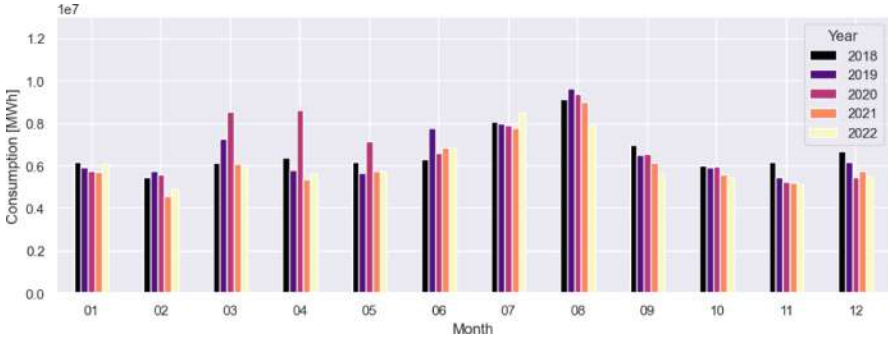


Fig. 4. Monthly consumption of the consumer cluster associated with S_3^* . The yearly seasonality is evident. The anomaly observed during the months of March, April and May 2020 correspond to the COVID-19 lockdown period.

believe these first results encourage us to address these limitations in our future research.

Acknowledgments. Koechlin and Secchi acknowledge the support by MUR, grant Dipartimento di Eccellenza 2023–2027. The authors want to thank Elena Degli Innocenti, Barbara Santini, Simona Vazio and Alessandro Venturi of the Statistics Group of Terna S.p.A for interesting and deep discussions on the topics developed in this paper.

References

1. Terna Developer API, Terna S.p.A. (2023). www.terna.it/en/electric-system/transparency-report/total-load
2. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics (2005). <https://doi.org/10.1007/b98888>
3. Zanini, P.: Blind source separation methods for high dimensional, massive, and complex data. Doctoral dissertation, Politecnico di Milano, Milan, Italy (2014). <https://hdl.handle.net/10589/89867>
4. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems 13 - Proceedings of the 2000 Conference, NIPS 2000 (2000). <https://doi.org/10.5555/3008751.3008829>
5. Gillis, N.: Nonnegative matrix factorization. Chapter 4, p. 99. Society for Industrial and Applied Mathematics (2020). <https://doi.org/10.1137/1.9781611976410>
6. Monthly Industrial Electricity Consumption Index (IMCEI). Terna S.p.A. <http://www.terna.it/en/electric-system/statistical-data-forecast/imcei>
7. Monthly Services Index (IMSER). Terna S.p.A. <https://www.terna.it/en/media/press-releases/detail/electricity-consumption-january-2024>



Forecasting Illiquidity Tail Risk and Its Economic Determinants

Demetrio Lacava^{1(✉)} and Paolo Santucci de Magistris²

¹ Department of Economics, University of Messina, Piazza Pugliatti, 1, 98122 Messina, Italy
dlacava@unime.it

² Department of Economics and Finance, Luiss University, Viale Romania 32,
00197 Roma, Italy

Abstract. We investigate the possibility of forecasting market illiquidity and assess the role of conditional jumps in determining the distributional properties of the realized Amihud with the goal of introducing the concept of *Illiquidity-at-Risk* (IlliquaR). We show that the inclusion of jumps drastically increases the forecasting capability of the models at least in a short horizon. Finally, accounting for illiquidity jumps allows for a correct probability coverage of extreme illiquidity events.

Keywords: Liquidity · Stochastic Volatility · Jumps · Forecast

1 Introduction

Beside volatility, a key ingredient for the efficient functioning of financial markets is liquidity. While, the theoretical properties of market liquidity were largely studied in the current literature, very little has been done as regards other aspects such as empirically analyzing the distributional properties of the realized Amihud. Taking the realized Amihud (Ranaldo and Santucci de Magistris 2022; Lacava et al., 2023) as a proxy for market illiquidity, we fill this gap in the literature through several alternative econometric specifications (both in a linear and non-linear context) which allow us to characterize the main dynamic and distributional properties of the realized Amihud (positive skewness and fat tails, in particular). To this end, we investigate the role played by illiquidity jumps (Caporin et al. 2017), by introducing the concept of Illiquidity at risk *IlliquaR*, i.e. the probability of observing abnormally high levels of illiquidity in a given day. Basing on the time series of the daily realized Amihud of several stock indices, we find that the inclusion of the jump term allows for a correct probability coverage of these illiquid events. Finally, we find that the probability of observing an IlliquaR violation is higher during downturn periods of the market and it is positively affected by VIX and risk aversion.

This paper illustrates the theoretical framework of our IlliquaR measure in Sect. 2; while the empirical application is discussed in Sect. 3. Finally, Sect. 4 concludes the paper with some remarks.

2 Illiquidity-at-Risk

We define IlliquaR - analogously to the Value-at-Risk (VaR) in the case of large negative returns - as the level, $\mathcal{A}(\alpha)$, that might be exceeded by illiquidity, A_t^* , with probability α ; in other words, the IlliquaR is a value such that

$$\Pr(A_t^* > \mathcal{A}(\alpha) | \mathcal{F}_{t-1}) = \alpha, \quad t = 1, 2, \dots, T, \quad (1)$$

where $\mathcal{F}_{t-1}$ is the information set up to time $t - 1$. The evaluation of the probability in (1) requires several ingredients. First, the true illiquidity process needs to be approximated by a precise measurement: we consider the *realized Amihud*, introduced by Rinaldo and Santucci de Magistris (2022), which is defined as the ratio between the realized power variation (Barndorff-Nielsen and Shephard, 2003) of order one and volume, both computed on a given unit interval (e.g. a day, a week, a month).¹ Second, a dynamic model for illiquidity must be developed with the goal of extracting information from $\mathcal{F}_{t-1}$ useful for predicting illiquidity at time t .² Third, a crucial ingredient for the computation of IlliquaR is the choice of the distribution for computing the conditional probability of tail events. Finally, the evaluation of our IlliquaR measure is based on the Berkowitz test (Berkowitz, 2001) for the adequacy of the density function with the realization of the dependent variable. Considering the flexibility of such a test, it can be applied to evaluate the fit of the full density as well as to specific quantile. In our application, we consider three quantile, corresponding to the 1, 5 and 10% of the Normal distribution. Importantly, the test is applied on the conditional density function (CDF) of A_t as expressed in Eq. (2)

$$y_t = F(A_t^* | \mathcal{F}_{t-1}) = \int_0^{A_t^*} f_A(A_t^* | \mathcal{F}_{t-1}) dA_t^*. \quad (2)$$

In details, for the HAR models, the CDF is that of a Normal distribution; for the MEM specifications, $F(A_t | \mathcal{F}_{t-1})$ is given by the Gamma CDF; finally, for the jump models, Eq. (2) is built as a mixture of a Gamma and a Kappa CDFs. The test is constructed under the null hypothesis that the model is correctly specified: if this is the case, the empirical CDF follows an uniform distribution, i.e. $y_t \sim U(0, 1)$. Basing on this consideration, the test is derived as a LR test for the comparison of the empirical CDF, y_t , and a new truncated variable, computed starting from the inverse of the standard normal distribution of y_t itself. In other words, with $s_t = \Phi^{-1}(y_t)$, we compute the new truncated variable, s_t^* as

$$s_t^* = \begin{cases} \mathcal{A}(\alpha) & \text{if } s_t \geq \mathcal{A}(\alpha) \\ s_t & \text{if } s_t < \mathcal{A}(\alpha). \end{cases} \quad (3)$$

¹ Lacava et al. (2023) show that realized Amihud converges to the true illiquidity signal aggregated on a given interval of unit length. In their framework, A^* denotes the *integrated illiquidity* that is defined as $A^* = \frac{1}{\int_0^1 \ell(s) ds}$, where $\ell(s)$ measures the instantaneous sensitivity of trades to new information, and it is a measure of a dimension of illiquidity related with market depth.

² We have considered several specifications: the linear HAR and the asymmetric HAR (AHAR), the MEM and the AMEM (with μ_t following either a GARCH(1,1) or a HAR process), and AMEM-J. For the latter, in addition to the AMEM(1,1)-J, we estimate several specifications: AMEM(1,2)-J, G-AMEM-HAR-J, MEM-HAR-J, MEM-AHAR-J.

The (right) tail coverage test can be derived using the LR principle based on the censored normal density of s_t^* . Under the null of correct tail coverage the test statistic is distributed as $\chi^2(2)$ as it corresponds to a test for comparing the mean and variance estimated from the truncated likelihood of s_t^* to those expected under correct coverage, i.e. zero mean and unit variance.

2.1 Assessing and Forecasting Illiquidity Tail Risk

Our empirical analysis focuses on the assessment of daily time series of the realized Amihud of the global stock market, as represented by four of the largest stock indexes: S&P500, FTSE100, DAX30, and Nikkei 225. Daily illiquidity is computed as the ratio between the daily realized volatility (RV, computed using returns sampled at 5-minutes intervals), and the daily volume.³ In particular, we assess the ability of each considered models (as listed in footnote 2) in providing a good characterization of illiquidity risk, that is the ability of providing a coverage of the probability of illiquidity spikes, as measured by IlliquaR.

The top panel of Table 1 reports the p -value of the Berkowitz test for the in-sample IlliquaR computed at the 1, 5 and 10% of the Gaussian distribution. It is evident how - regardless of the specification for the conditional mean μ_t - the jump models outperform the models without jump for the characterization of IlliquaR. In general, for all the considered quantiles, the jump models seem to be appropriate for the correct specification of the upper tail, as we can never reject the null of correct specification, i.e. the p -value is always well above the threshold of 0.1 (with the exceptions of the 5% quantile of the normal distribution for the S&P 500 and FTSE 100 index). Once again, we interpret this result as evidence supporting the view that realized illiquidity requires to be modeled through a model that assigns the correct probability to extreme realizations, those lying on the right tail of the distribution. Furthermore, the Berkowitz test on the out-of-sample IlliquaR corroborates the intuition that jumps are crucial to assign the correct probability to extreme illiquidity events. In details, as reported in bottom panel of Table 1, we always reject the null hypothesis of correct tail coverage for no-jump models; in contrast, when jumps are included, in most of the case, the associated p -value is at least higher than 0.01, while it is higher than the 0.1 threshold for the 1% quantile of the normal distribution for FTSE 100 and DAX 30. In conclusion, even in the out-of-sample framework, the inclusion of jumps is useful to assign the correct probability to extreme realizations.

3 Illiquidity Tail Risk and Its Economic Determinants

Finally, with the aim of shedding some light on the determinants of illiquidity we have investigated whether the probability of the IlliquaR violations, defined as $\zeta_t = \max(\text{Illiqu}_t - \mathcal{J}_t(\alpha), 0)$, is affected by some exogenous variables. In doing so, we have

³ Given the high correlation between RPV and RV, replacing RPV with RV is virtually irrelevant. Data on RV is provided by the Oxford Man Institute, while data on daily volume is provided by Datastream.

Table 1. Table reports the p-values of the Berkowitz test for the insample IlliquAR at 1, 5 and 10% quantiles of the Gaussian distribution for realized Amihud.

Full Sample															
S&P 500							FTSE 100								
No-Jump Models				Jump Models			No-Jump Models				Jump Models				
	1%	5%	10%		1%	5%	10%		1%	5%	10%		1%	5%	10%
HAR	0.0000	0.0000	0.0000	MEMJ	0.1826	0.4070	0.2392	HAR	0.0000	0.0000	0.0000	MEMJ	0.6404	0.2706	0.3388
AHAR	0.0000	0.0000	0.0000	AMEMJ	0.4627	0.0073	0.7580	AHAR	0.0000	0.0000	0.0000	AMEMJ	0.3940	0.4534	0.0016
MEM	0.0000	0.0000	0.0000	AMEMJ(2,1)	0.4708	0.0017	0.4868	MEM	0.0000	0.0000	0.0000	AMEMJ(2,1)	0.1835	0.0255	0.8520
AMEM	0.0000	0.0000	0.0000	G-AMEM-HAR-J	0.3896	0.0183	0.4171	AMEM	0.0000	0.0000	0.0000	G-AMEM-HAR-J	0.0996	0.0369	0.8519
MEM HAR	0.0000	0.0000	0.0000	MEMJ HAR	0.1133	0.0113	0.0494	MEM HAR	0.0000	0.0000	0.0000	MEMJ HAR	0.6259	0.0215	0.2992
MEM AHAR	0.0000	0.0000	0.0000	MEMJ AHAR	0.0406	0.0038	0.1455	MEM AHAR	0.0000	0.0000	0.0000	MEMJ AHAR	0.0912	0.0059	0.6727
G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000			
DAX 30							N225								
No-Jump Models				Jump Models			No-Jump Models				Jump Models				
	1%	5%	10%		1%	5%	10%		1%	5%	10%		1%	5%	10%
HAR	0.0023	0.0000	0.0000	MEMJ	0.3233	0.3926	0.2689	HAR	0.0000	0.0000	0.0000	MEMJ	0.1416	0.7216	0.9220
AHAR	0.0106	0.0000	0.0000	AMEMJ	0.3185	0.6201	0.2370	AHAR	0.0000	0.0000	0.0000	AMEMJ	0.4069	0.9422	0.9570
MEM	0.0000	0.0000	0.0000	AMEMJ(2,1)	0.0640	0.7421	0.04370	MEM	0.0000	0.0000	0.0000	AMEMJ(2,1)	0.5675	0.8925	0.7467
AMEM	0.0000	0.0000	0.0000	G-AMEM-HAR-J	0.2387	0.7150	0.1254	AMEM	0.0000	0.0000	0.0000	G-AMEM-HAR-J	0.2962	0.7084	0.9380
MEM HAR	0.0000	0.0000	0.0000	MEMJ HAR	0.1706	0.6213	0.3469	MEM HAR	0.0000	0.0000	0.0000	MEMJ HAR	0.6327	0.8174	0.9787
MEM AHAR	0.0000	0.0000	0.0000	MEMJ AHAR	0.1540	0.3359	0.2482	MEM AHAR	0.0000	0.0000	0.0000	MEMJ AHAR	0.4467	0.9934	0.8515
G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000			
Out-of-Sample															
S&P 500							FTSE 100								
No-Jump Models				Jump Models			No-Jump Models				Jump Models				
	1%	5%	10%		1%	5%	10%		1%	5%	10%		1%	5%	10%
HAR	0.0000	0.0000	0.0000	MEMJ	0.0095	0.0010	0.0012	HAR	0.0000	0.0000	0.0000	MEMJ	0.1049	0.0189	0.0680
AHAR	0.0000	0.0000	0.0000	AMEMJ	0.0155	0.0001	0.0005	AHAR	0.0000	0.0000	0.0000	AMEMJ	0.0914	0.0015	0.0901
MEM	0.0462	0.0001	0.0000	AMEMJ(2,1)	0.0160	0.0001	0.0008	MEM	0.0000	0.0000	0.0000	AMEMJ(2,1)	0.1467	0.0009	0.0893
AMEM	0.0118	0.0000	0.0000	G-AMEM-HAR-J	0.0170	0.0012	0.0111	AMEM	0.0000	0.0000	0.0000	G-AMEM-HAR-J	0.0921	0.0002	0.0262
MEM HAR	0.0033	0.0008	0.0000	MEMJ HAR	0.0090	0.0033	0.0073	MEM HAR	0.0000	0.0000	0.0000	MEMJ HAR	0.6243	0.0019	0.0060
MEM AHAR	0.0013	0.0000	0.0000	MEMJ AHAR	0.0031	0.0025	0.0199	MEM AHAR	0.0000	0.0000	0.0000	MEMJ AHAR	0.3446	0.0003	0.0226
G-AMEM-HAR 0.0004 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000			
DAX 30							N225								
No-Jump Models				Jump Models			No-Jump Models				Jump Models				
	1%	5%	10%		1%	5%	10%		1%	5%	10%		1%	5%	10%
HAR	0.0001	0.0000	0.0000	MEMJ	0.7276	0.0066	0.0049	HAR	0.0001	0.0000	0.0000	MEMJ	0.0001	0.0001	0.0001
AHAR	0.0003	0.0000	0.0000	AMEMJ	0.6767	0.0031	0.0012	AHAR	0.0001	0.0000	0.0000	AMEMJ	0.0000	0.0000	0.0000
MEM	0.0007	0.0307	0.0878	AMEMJ(2,1)	0.7064	0.0036	0.0044	MEM	0.0000	0.0000	0.0000	AMEMJ(2,1)	0.0000	0.0000	0.0000
AMEM	0.0015	0.0379	0.0596	G-AMEM-HAR-J	0.5787	0.0341	0.0086	AMEM	0.0000	0.0000	0.0000	G-AMEM-HAR-J	0.0000	0.0000	0.0000
MEM HAR	0.0011	0.0008	0.0206	MEMJ HAR	0.3162	0.0811	0.0076	MEM HAR	0.0000	0.0000	0.0000	MEMJ HAR	0.0001	0.0000	0.0000
MEM AHAR	0.0001	0.0026	0.0202	MEMJ AHAR	0.6544	0.0167	0.0019	MEM AHAR	0.0000	0.0000	0.0000	MEMJ AHAR	0.0000	0.0000	0.0000
G-AMEM-HAR 0.0005 0.0029 0.0397				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000				G-AMEM-HAR 0.0000 0.0000 0.0000			

estimated a logit model for the limited dependent variable I_t , which is equal to 1 if $\text{Illiq}_t > \mathcal{I}_t(\alpha)$, and 0 otherwise. In practice, we have estimated the following model

$$Pr(I_t = 1|X) = \frac{\exp(\beta\mathbf{X})}{1 + \exp(\beta\mathbf{X})}, \quad (4)$$

where $\mathbf{X}$ is the matrix of regressors and β is the vector of coefficients. As regressors, we consider a set of predetermined macro-financial variables that represent the level of uncertainty in the economy and in the financial markets. These are: VIX, i.e. the popular Cboe Volatility Index that is based on the S&P 500 index and representing the option-implied expected volatility; EPU, i.e. the Economic Policy Uncertainty Index by Baker et al. (2016); TED that is the yield spread between the US three-month Libor and

T-bills, representing a measure of funding strains; finally, the variable D is a dummy variable for negative returns. Furthermore, by following the procedure developed by Bekaert et al. (2013), it is possible to further decompose the VIX index in order to account for both market uncertainty and risk aversion.

Table 2 summarizes estimation results for the full sample (in columns 1–2) and out-of-sample analysis (columns 3–4), where the regressors are considered at time t . In such

Table 2. Estimation results from Logit model. Robust standard errors in parenthesis. Average marginal effects in square brackets. Coefficients denoted by superscript * are multiplied by 100. The superscripts a, b and, c denote significant coefficients at 1%, 5% and, 10% level, respectively.

	Full sample		Out of sample	
	Mod.1	Mod.2	Mod.1	Mod.2
Parameter Estimates				
Constant	−4.0424 ^a (0.1729)	−3.1838 ^a (0.2774)	−5.0866 ^a (0.4397)	−3.1685 ^a (0.5903)
VIX_t	0.0131 ^b (0.0065)		0.1116 ^a (0.0175)	
	[0.0005]		[0.0015]	
$Uncertainty_t$		−152.956 ^a (52.0721)		−99.5196 (89.7760)
		[−6.1645]		[−1.3742]
$RiskAversion_t$		0.0012 ^a (0.0004)		0.0021 ^a (0.0008)
		[0.0001]		[0.0000]
EPU_t	0.0013 ^c (0.0008)	0.0022 ^a (0.0007)	−0.0048 ^b (0.0021)	−0.0029 (0.0020)
	[0.0001]	[0.0001]	[−0.0001]	[0.0000]
TED_t	0.0024 (0.0170)	0.0025 (0.0140)	−2.9462 ^a (0.7220)	−3.4549 ^a (0.8577)
	[0.0001]	[0.0001]	[−0.0391]	[−0.0477]
D_t	1.1996 ^a (0.1560)	1.1104 ^a (0.1551)	1.5621 ^a (0.3710)	1.5431 ^a (0.3783)
	[0.0544]	[0.0487]	[0.0245]	[0.0250]
Model Fit				
Rate of success	94.80%	94.90%	96.40%	96.50%
LR test	80.5355 (0.0000)	97.0811 (0.0000)	77.4052 (0.0000)	68.6476 (0.0000)

a case, the dummy variable is significant with a higher marginal effect around 0.05; also the VIX is significant at the highest level and it is also crucial to forecast the probability of IlliquaR violations in an out of sample context (column 3). The uncertainty coefficient is negative while risk aversion is positive and significant both for the full sample (column 2) and the out-of-sample analysis (column 4): this gives crucial information to investors, who may anticipate a peak in tail risk, exceeding a given threshold level $\mathcal{J}_t(\alpha)$, by updating their information set with the forecast of some observable variables such as the VIX, uncertainty and, risk aversion.

4 Conclusions

In this paper we investigate the contribution of illiquidity jumps in determining the distributional properties of the realized Amihud. Our empirical analysis reveals that jumps are a significant component of the illiquidity risks. Assigning the correct probability to extreme realizations of the realized Amihud is of fundamental importance, especially for forecasting purposes (at least in a short horizon, 1-step-ahead). The Berkowitz test, in particular, highlights the importance on including jumps in modelling illiquidity, in order to assigns the correct probability to tail events. Finally, it emerges that the probability of observing an IlliquaR violation is higher during downturn periods of the market - when returns are negative - and it is positively affected by risk aversion.

References

- Baker, S.R., Bloom, N., Davis, S.J.: Measuring economic policy uncertainty. *Q. J. Econ.* **131**(4), 1593–1636 (2016)
- Barndorff-Nielsen, O.E., Shephard, N.: Realized power variation and stochastic volatility models. *Bernoulli* **9**(2), 243–265 (2003)
- Bekaert, G., Hoerova, M., Duca, M.L.: Risk, uncertainty and monetary policy. *J. Monet. Econ.* **60**, 771–778 (2013)
- Berkowitz, J.: Testing density forecasts, with applications to risk management. *J. Bus. Econ. Stat.* **19**(4), 465–474 (2001)
- Caporin, M., Rossi, E., Santucci de Magistris, P.: Chasing volatility: a persistent multiplicative error model with jumps. *J. Econometr.* **198**(1), 122–145 (2017)
- Lacava, D., Ranaldo, A., Santucci de Magistris, P.: Realized illiquidity. In: Swiss Finance Institute Research Paper, pp. 22–90 (2023)
- Ranaldo, A., Santucci de Magistris, P.: Liquidity in the global currency market. *J. Financ. Econ.* **146**, 859–883 (2022)



Statistical Modelling of Population Gaps in Mortality: The Skellam Distribution to Model Differences in Death Counts

Giacomo Lanfiuti Baldi¹(✉) and Andrea Nigri²

¹ Department of Statistics, Sapienza University of Rome, Rome, Italy
`giacomo.lanfiutibaldi@uniroma1.it`

² Department of Economics, Management and Territory, University of Foggia,
Foggia, Italy
`andrea.nigri@unifg.it`

Abstract. Understanding and modeling mortality patterns, particularly differences in death between populations, is crucial for demographic analysis and public health planning. In this study, we compare three statistical models in the Age-Period context to analyse differences in death counts. The models considered are based on Double Poisson, Bivariate Poisson, and Skellam distributions, each offering distinct advantages in capturing the underlying mortality trends. Using mortality data spanning from 2003 to 2020 related to the two major causes of death in Italy: cancer and diseases of the circulatory system, our analysis reveals significant temporal and age-related variations in the mortality patterns of the two causes. We find that the Skellam distribution model, while a novel approach in this field, demonstrates superior accuracy and simplicity in capturing mortality differentials. Our findings underscore the potential utility of the Skellam distribution in mortality gaps analysis.

Keywords: Mortality modeling · Causes-of-deaths · Skellam regression · Age-Period data

1 Introduction

Modeling mortality stands as a fundamental tool in understanding demographic patterns and projecting future trends. Mortality data are often reported as count data within specific age-period domains for various populations. The simplest models for mortality patterns typically involve age and period as the sole explicative variables, since we do not need any more information than what is intrinsically contained in the data. This approach forms the cornerstone for exploring and comprehending the intricate dynamics of mortality across diverse populations. Furthermore, our focus lies in delineating the disparities between two populations—these populations could manifest as two genders, distinct countries, or differing causes of death. We aim to provide a statistical framework for jointly modelling the mortality of two populations, beginning with the shared

assumption that the count of deaths at each age and year follows a Poisson distribution [1].

Given the mortality data for two populations, denoted as A and B, across ages $(1, \dots, X)$ and years $(1, \dots, T)$, our aim is to determine the average period trend and age-specific profile of the differences in mortality. This research helps to build a robust model capable of analysing and forecasting mortality disparities between the two populations. The fundamental assumption underlying our approach is that deaths at age x and year t for both populations (D_{xt}^A, D_{xt}^B) are generated by a Poisson process with different intensity parameters λ_{xt}^A and λ_{xt}^B :

$$D_{xt}^A \sim \text{Pois}(\lambda_{xt}^A), \quad D_{xt}^B \sim \text{Pois}(\lambda_{xt}^B). \quad (1)$$

We define the gap in deaths between the two populations at age x and year t as:

$$G_{xt} = D_{xt}^A - D_{xt}^B \quad (2)$$

2 Models

In this paper, we compare three possible modelling for Age-Period mortality data.

Double Poisson. The simplest model is to use two independent Poisson variables where the relevant parameters are constructed as the sum of ages (β_x^α) and periods (β_x^π) effects. We define this model as *Double Poisson*. It considers two different models on the two Poisson's parameters:

$$\begin{aligned} \log(\lambda_{xt}^A) &= \mu^A + \beta_x^{A,\alpha} + \beta_x^{A,\pi} \\ \log(\lambda_{xt}^B) &= \mu^B + \beta_x^{B,\alpha} + \beta_x^{B,\pi} \end{aligned} \quad (3)$$

If the interest is in estimating the differences between the deaths of the two populations per year and age, it's possible to calculate:

$$\widehat{G_{xt}} = \widehat{\lambda_{xt}^A} - \widehat{\lambda_{xt}^B} \quad (4)$$

Bivariate Poisson. The Double Poisson model is as simple as it is simplistic. By considering processes generated by two independent Poissons it is making a partially unrealistic assumption. If we consider two subpopulations of a single population (e.g. two sexes or two causes of death) or two populations experiencing similar environmental and exogenous phenomena (e.g. two neighbouring countries), the correlation between the deaths of the two can be high.

A different approach from the Double Poisson model involves considering a *Bivariate Poisson* distribution for the two outcome variables. In this model, the marginal distributions remain simple Poisson distributions, but the random variables are now dependent [2].

The random variables $X = X_1 + X_3$ and $Y = X_2 + X_3$ follow jointly a Bivariate Poisson distribution:

$$(X, Y) \sim \text{BP}(\lambda_1, \lambda_2, \lambda_3)$$

Marginally each random variable follows a Poisson distribution:

$$X \sim \text{Pois}(\lambda_1 + \lambda_3), \quad Y \sim \text{Pois}(\lambda_2 + \lambda_3)$$

Moreover, $\text{Cov}(X, Y) = \lambda_3$ and this means that λ_3 is a measure of dependence between the two random variables. It is easy to note that if $\lambda_3 = 0$, then the Bivariate Poisson corresponds to a Double Poisson.

In the case of the deaths count of two populations we therefore have:

$$(D_{xt}^A, D_{xt}^B) \sim \text{BP}(\lambda_{xt}^A, \lambda_{xt}^B, \lambda_{xt}^{AB}) \quad (5)$$

In this modelling based on the Bivariate Poisson distribution, we consider three different models on the three parameters:

$$\begin{aligned} \log(\lambda_{xt}^A) &= \mu^A + \beta_x^{A,\alpha} + \beta_x^{A,\pi} \\ \log(\lambda_{xt}^B) &= \mu^B + \beta_x^{B,\alpha} + \beta_x^{B,\pi} \\ \log(\lambda_{xt}^{AB}) &= \mu^{AB} \end{aligned} \quad (6)$$

Following [2] we consider models with constant λ_{xt}^{AB} (no covariates on λ_{xt}^{AB}) because such models are easier to interpret.

In estimating the gap, we have simply:

$$\widehat{G_{xt}} = \widehat{\lambda_{xt}^A} + \widehat{\lambda_{xt}^{AB}} - \widehat{\lambda_{xt}^B} + \widehat{\lambda_{xt}^{AB}} = \widehat{\lambda_{xt}^A} - \widehat{\lambda_{xt}^B}. \quad (7)$$

Skellam. By using the Bivariate Poisson distribution in the estimation of the gap between deaths the correlation component disappears. We therefore propose the use of the *Skellam* (or *Poisson Difference*) distribution to directly model the gap instead of jointly modelling the number of deaths of the two populations.

The Poisson Difference distribution was introduced by Skellam in 1946 [3]. Consider two discrete random variables X and Y and their difference $Z = X - Y$. The probability function of the difference Z is a discrete distribution defined on the set of integer numbers (positive and negative) $\mathbb{Z}$. This is true for any pair of random variables (X, Y) that can be written as $X = W_1 + W_3$ and $Y = W_2 + W_3$, with $W_1 \sim \text{Pois}(\lambda_1)$ and $W_2 \sim \text{Pois}(\lambda_2)$ and W_3 following any distribution. The joint distribution of variables X and Y forms a Bivariate distribution, where the correlation arises from the shared stochastic element present in both variables, denoted as W_3 [4]. The mean of a Skellam distribution is given by $\mathbb{E} = \lambda_1 - \lambda_2$. We assume that $G_{xt} \sim \text{PD}(\lambda_{xt}^C, \lambda_{xt}^D)$ and just like before we model it using Age and Period as covariates:

$$\begin{aligned} \log(\lambda_{xt}^C) &= \mu^C + \beta_x^{C,\alpha} + \beta_x^{C,\pi} \\ \log(\lambda_{xt}^D) &= \mu^D + \beta_x^{D,\alpha} + \beta_x^{D,\pi} \end{aligned} \quad (8)$$

The fitted values of the regression directly provide estimates of the death gap:

$$\widehat{G_{xt}} = \widehat{\lambda_{xt}^C} - \widehat{\lambda_{xt}^D} \quad (9)$$

The use of the Skellam distribution only requires the series of differences in the deaths of populations A and B and not the two death series themselves.

Model sets such as those listed here are often used in the context of football score modelling ([4–6]), but the use of Skellam in the context of mortality modelling is a novelty.

3 Results

To test and compare the proposed modelling, we present the study of the differences between deaths due to the two main causes of death in Italy: cancer and diseases of the circulatory system.

Deaths before the age of 50 are few for both causes and the differences are close to 0 (Fig. 1). Between the ages of 50 and around 75–80, more deaths are due to cancer, and at later ages more deaths are due to diseases of the circulatory system.

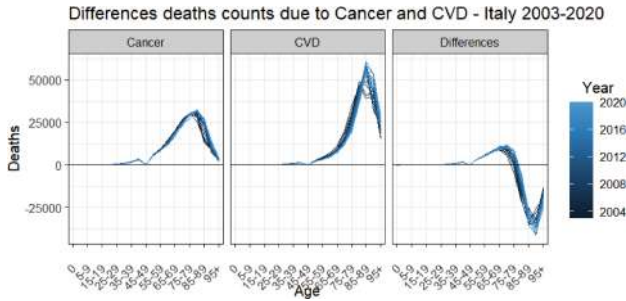


Fig. 1. Trends by age and year in Cancer and CVD death counts and their differences. In Italy between 2003 and 2020, both sexes combined.

We applied the three models explained in Sect. 2, considering age and years as categorical variables with the first category treated as references (age 0 and the year 2003). Thus, we obtained two sets of age and period effects for each model. The effects are reported in Fig. 2.

For Double Poisson and Bivariate Poisson, the estimated effects represent the trend of the two causes of death over time and ages, whereas for Skellam this is not true. For this last modelling, we obtain the pattern of effects that jointly best approximate the observed differences.

Overall, we observe an increase in differences over time, mainly in the ages between 70 and 90. In addition, modelling using Bivariate Poisson provides an estimate of the covariance parameter between the two causes, which turns out to be 2.78.

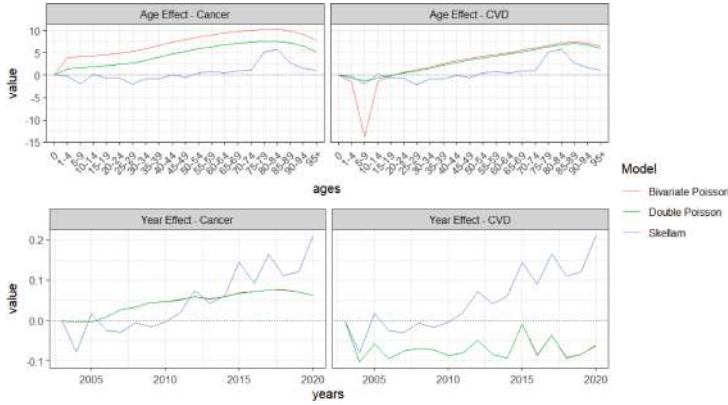


Fig. 2. Age-Period of the two causes-of-deaths and relative differences

We compare the accuracy of the three proposed models in estimating the differences in death counts (G_{xt}). To evaluate the accuracy we use the Root-Mean-Square Error (RMSE) and Mean Absolute Error (MAE). The values of the metrics for the three models are reported in Table 3.

Model	RMSE	MAE
Skellam	1476.25	724.40
Bivariate Poisson	1819.11	857.82
Double Poisson	1820.21	858.33

Although the Skellam-based model needs less information (not the two death series, but only the observed differences), it appears to be the model that best fits the data. The models based on Double Poisson and Bivariate Poisson have similar performance, the model that considers the correlation between the death series is only slightly better.

4 Conclusion

In this paper, we propose three different modelling approaches for the study of death counts and their differences. The three modelling starts from the typical assumption that deaths follow a Poisson distribution at each age and year. We then defined three Age-Period models for differences between counts based on: Double Poisson, Bivariate Poisson and Skellam distribution. These three models represent, respectively: a classical approach, a more sophisticated approach and a novelty in the study of mortality gaps.

We have chosen as a case study for the application of the three models, the analysis of the differences in deaths due to the two major causes of death in

Italy: cancer and diseases of the circulatory system. Working with causes of death allows us to avoid considering the number of risk exposures in the modelling, but we can still compare the two death series. This is because the exposure populations for the two causes are the same and the differences in mortality rates and death counts are the same.

Comparing the fitting of the three models directly on the differences between the death counts, we observed that the model based on the Skellam distribution, a novelty in this field, is by far the best as well as the one requiring the least amount of input information.

In light of these results, the Skellam distribution could be useful for jointly and consistently modelling two causes of death. Moreover, it is easy to obtain a generalisation for joint modelling of deaths of any two populations (the two sexes, two countries or two different economic conditions). By exploiting the Skellam distribution, it will also be possible to define coherent models for forecasting.

References

1. Brillinger, D.R.: A biometrics invited paper with discussion: the natural variability of vital rates and associated statistics. In: *Biometrics*, pp. 693–734 (1986)
2. Karlis, D., Ntzoufras, I.: Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *J. Stat. Softw.* **14**, 1–36 (2005)
3. Skellam, J.G.: The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. Royal Stat. Soc. Ser. A: Stat. Soc.* **109**(3), 296–296 (1946)
4. Karlis, D., Ntzoufras, I.: Bayesian modelling of football outcomes: using the Skellam’s distribution for the goal difference. *IMA J. Manage. Math.* **20**(2), 133–145 (2009)
5. Egidi, L., Pauli, F., Torelli, N.: Combining historical data and bookmakers’ odds in modelling football scores. *Stat. Model.* **18**(5–6), 436–459 (2018)
6. Baio, G., Blangiardo, M.: Bayesian hierarchical model for the prediction of football results. *J. Appl. Stat.* **37**(2), 253–264 (2010)



The Challenge of Microdata Dissemination. A Methodology for a Public Use File Version of the WHIP Database

Roberto Leombruni¹, Consuelo Rubina Nava^{2(✉)}, and Marina Zanatta¹

¹ Department of Economics and Statistics, University of Turin, Turin, Italy
{roberto.leombruni,marina.zanatta}@unito.it

² Department of Economics and Political Science, University of Aosta Valley,
Aosta, Italy
c.nava@univda.it

<https://www.est.unito.it/persona/roberto.leombruni/>,
<https://www.univda.it/docenti/nava-consuelo-rubina/>

Abstract. This study proposes a methodology for creating a Public Use File version of the WHIP database. The study aims to implement statistical methods to estimate the risk of identifying a subject belonging to the sample. The 3-step methodology starts with a mock interview of a subsample of the data, proceeds with the estimation of the identification risk, and applies standard anonymization methods, such as global recoding and local suppression, to the units at risk of re-identification.

Keywords: Anonymization · Public Use File - PUF · Global recode · Key variables · k-anonymity · WHIP (Work History Italian Panel) database

1 Introduction

Data producers must respect, when releasing microdata, the principles of protection of confidentiality enshrined in the current legislation in force [1], recently redefined by the European General Data Protection Regulation 2016/679 (GDPR), transposed into Italian law decree (LD) 24/2023.

This contribution presents the methodology for designing a Public Use Files (PUF) version of WHIP (Work History Italian Panel). The latter is an extensive database of work and health biographies developed within the Italian National Statistical Programme (PSN). The methodology is based on three steps: 1) a mock interview of a subsample of the original data; 2) the estimation of the risk of identifying a subject belonging to the sample; 3) the application of standard techniques of anonymization to units at risk of identification [2]. To this aim, variables that an intruder could employ to recognize a statistical unit, the key variables [3], are identified.

This paper is structured as follows. Sections 2 and 3 explain the methodology and the data, respectively. The anonymization results are illustrated in Sect. 4 while Sect. 5 concludes the paper by envisaging further research and highlighting the main limitations.

2 Methodology

To carry out the second methodological step, global risk measures of re-identification [4,5] must be introduced. Let F_k and f_k be the number of individuals in the population of size N and in a sample, where $k = 1, \dots, K$ denotes all possible combinations of the key variables modes. Firstly, τ_1 counts how many uniques in the sample ($f_k = 1$) are also uniques in the population ($F_k = 1$), namely

$$\tau_1 = \sum_{k=1}^K I(F_k = 1 \cap f_k = 1).$$

Secondly, τ_2 is the ratio between the number of unique samples and the frequency of observation of the k combination in the population, i.e.

$$\tau_2 = \sum_{k=1}^K I(f_k = 1) \frac{1}{F_k}.$$

Generally, the population is considered a sample of the so-called superpopulation thus F_k is unknown and the following prior could be assigned to it $F_k \sim \text{Poisson}(\lambda_k)$ $f_k \sim \text{Poisson}(\omega \lambda_k)$. Considering the following Bayesian hierarchy [6], the relationship between superpopulation, population, and sample allows us to model the conditional probability mass function $F_k|f_k$ as follows:

$$\begin{aligned} \pi_k &\sim [\pi_k] \propto \frac{1}{\pi_k} \quad \forall k = 1, \dots, K \\ F_k|\pi_k &\sim \text{Poisson}(N^* \pi_k), \quad F_k = 0, 1, \dots, \\ f_k|F_k, \pi_k, p_k &\sim \text{Binomial}(F_k, p_k), \quad f_k = 0, 1, \dots, F_k, \end{aligned}$$

where the size of the superpopulation N^* is assumed to be known and π_k represents the inclusion probability

$$F_k|f_k \sim f_k + \text{Poisson}(N^*(1 - \pi_k)).$$

$F_k|f_k$ is modeled with a negative binomial distribution with success probability p_k and number of successes f_k . In this setting, the individual risk of re-identification, r_k , is the posterior mean of $\frac{1}{F_k}$ concerning the distribution of $F_k|f_k$: $r_k = E\left(\frac{1}{F_k} | f_k\right) = \sum_{h > f_k} \frac{1}{h} P(F_k = h | f_k)$. Within this setting, the sample-specific measure used to understand a database's security level is k -anonymity. An observed unit violates k -anonymity if the frequency in the sample, given the l -th combination of key modes, is less than the threshold k [7]. The number of units that violate the k -anonymity, i.e.,

$$\sum_{i=1}^n I(f_l < k) i$$

is the risk measure adopted to classify a dataset as protected. Since we are dealing with a sample, a low level of k -anonymity (i.e., a number of units in

the sample with the same combination of key variable modalities higher than k) does not necessarily imply a re-identification risk. Thus, contributions in [6] identify units with a significant disclosure risk.

A set of anonymization techniques is subsequently applied precisely concerning these disclosed units. Data protection/anonymization methodologies considered here are summarized in Table 1. The use of specific methods corresponds to the different nature of the anonymized variable. For instance, categorical variables are treated, taking advantage of local suppression, post-randomization, rank swapping, and global recoding techniques. All these approaches, integrated with additive noise and microaggregation, are suitable for anonymizing numerical data.

Table 1. Anonymization methodologies

Methodology	Variable types	Advantages	Disadvantages
Noise addition [8]	Quantitative	Lower risk of successful attacks; individual data; complementary measure	Addition of inconsistent noise; less reliable data; needs a correct noise attribution
Swapping [9]	Qualitative quantitative	Variable modalities distribution unchanged; no estimates distortion; still identifiable individual's data; complementary measure	Choice of permutation attribute; deteriorated data; random permutation; less reliable data
Global recode [10]	Qualitative quantitative	Easy implementation; automated or guided; widely used for PUF; lower risk of successful attacks	No longer identifiable individual's data
Post randomization [11]	Qualitative quantitative	Individual data; correct transition probabilities	Estimates distortion; less reliable data
Local suppression [12]	Qualitative quantitative	No individual information and variable deletion	Estimates distortion
Micro aggregation [13]	Quantitative	Individual data; easily implementable algorithms; accurate aggregation criterion; correct definition of observations per group;	Less reliable data; alteration of variable levels; trade-off between techniques with fixed and variable numerosity; not widely used in PUF's building

3 Data

The WHIP database is based on the linkage of Italian administrative data coming from INPS (National Institute of Social Security), INAIL (National Compensation Authority), ISTAT (National Institute of Statistics), Ministry of Health,

and Ministry of Welfare. The reference population is composed of all individuals who worked or have worked in Italy over the period from 1985 to 2019. WHIP is extracted with a systematic sampling by birthday to ensure a representative population sample with an approximately 1:15 inclusion probability [14].

The level of detail in the original data is exceptionally high (the 3-anonymity evaluated at sections of the data is violated in 99.99%). In this way, we selected about 175,000 individuals, who were mocked-interviewed with 45 questions grouped in four thematic areas. The first draws the main characteristics of the individual; the second provides a retrospective analysis of the subject's work history; the third is focused on his/her employment situation at the time of the interview; and the fourth collects highly general information on the subject's health.

Among the information collected, we considered key variables, i.e., characteristics that can potentially allow the indirect identification of individuals, the following ones: sex; birth area; date of birth; has worked before; average income; start of current job; place of work; skill level; part-time; contract type; firm size; sector of activity; total income; transfers; number of accidents; maternity leave; and layoffs (see Table 2). The global recoding of the key variables has been calibrated to achieve a level of 3-anonymity comparable to the one we measured on the Italian Labour Force Survey MFR, using a similar set of key variables.

Table 2. Variable recoding specifications

Key variable	Original mode	Recoding
Sex	M-F	
Birth area	Provinces	NE, NW, C, S, Islands, Foreign
Date of birth	dd/mm/yyyy	10 years brackets
Has worked before	0, 1	0, 1
Average income	Numeric continuous	Quintiles
Start of current job	dd/mm/yyyy	1, 2
Place of work	Provinces	NE, NW, Center, South-Islands
Skill level	1, 2, 3, 4, 5	
Part-time	Numeric discrete	0, 1
Contract type	60 modes	3 modes
Firm size	Numeric discrete	3 modes
Sector of activity	Ateco 2007 Sub-Sections	Ateco 2007 Sections
Total income	Numeric continuous	Deciles
Transfers	Numeric continuous	0, 1
Number of accidents	Numeric continuous	0, 1 or more than 1
Maternity leave	0, 1	0, 1
Layoff	0, 1	0, 1

4 Results

Before conducting an anonymization analysis and individual risk estimation, the observations with non-structural missing values are eliminated. Based on the aforementioned key variables, the observations that present missing values are subject to a first temporary elimination operation (as they will be explicitly treated later). The subjects with missing observations, not attributable to a situation of absence of work at the time of the interview, are identified: those who were not working are identified, while missing people not associated with these conditions are eliminated. This involves losing 63,911 observations while obtaining a PUF with 111,469 respondents. The latter, with the recoded variables as in Table 2, have been studied to evaluate the current level of k -anonymity, which resulted in 39,287 and 53,349 observations violating the 2- and 3-anonymity, respectively. Following [6], we then estimated the disclosure risk on the units violating k -anonymity, comparing it with the initial level of risk before global recoding (Fig. 1), and applied anonymization techniques to the units exceeding a 1% level of risk. Micro-aggregation did not produce satisfactory results; thus, local suppression actions were undertaken. Two different local removal approaches can be carried out. First, attributing a missing value for the modalities of one or more key variables characterizing an observation and implying a high risk of identification. Second, it is possible to identify the individual risk for a single observation contained in the dataset according to [6] model and, taking into account the proportion with which sampling is carried out from the population of workers, to have a criterion with which to carry out local suppression. The individual risk estimate will be compared with a threshold level representative of the individual risk level above which to carry out local suppression. We can start by identifying the keys to begin carrying out local suppression and immediately verify the level of protection of the dataset achieved thanks to the action of local suppression on the single key. In this way, with a re-identification risk estimation model and the study of the unique samples, the anonymization of the file can be appropriately completed.

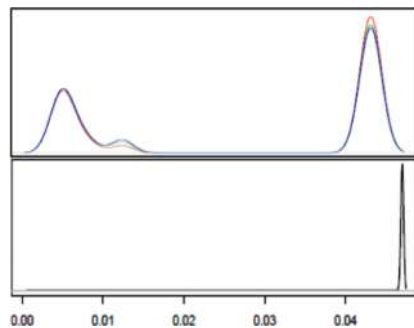


Fig. 1. Density of the identification risk after (top) and before (bottom) global recoding

5 Conclusion

This study presents a methodology for producing an anonymized version of the WHIP database, which will facilitate the dissemination of microdata on work and health biographies for research purposes. A mock interview and subsampling of the original data [15], plus global recoding of key variables, have proved to be a practical and easily implementable anonymization method. Given that the re-identification risk depends on the choice of key variables (numerosity and quality) and to the threshold level k , further research is needed, and other nonparametric Bayesian approaches to estimate disclosure risk [5] can be investigated.

References

1. Rubinstein, I.S., Hartzog, W.: Anonymization and risk. *Wash. L. Rev.* **91**, 703 (2016)
2. Gruppo di Lavoro Art. 29, Parere 5/2014 sulle tecniche di anonimizzazione, 0829/04/IT WP216
3. Mackey, E.: A best practice approach to anonymization. *Handbook of Research Ethics and Scientific Integrity*, pp. 323–343 (2020)
4. Rinott, Y., Shlomo, N.: A generalized negative binomial smoothing model for sample disclosure risk estimation. In: *International Conference on Privacy in Statistical Databases*, pp. 82–93. Springer, Berlin (2006). https://doi.org/10.1007/11930242_8
5. Carota, C., Filippone, M., Leombruni, R., Polettini, S.: Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models (2015)
6. Benedetti, R., Franconi, L.: Statistical and technological solutions for controlled data dissemination. In: *Pre-proceedings of New Techniques and Technologies for Statistics*, vol. 1, pp. 225–232. Sorrento (1998)
7. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k -anonymization. In: *21st International Conference on Data Engineering*, pp. 217–228 (2005)
8. Brand, R.: Microdata protection through noise addition. In: Domingo-Ferrer, J. (ed.) *Inference Control in Statistical Databases*, pp. 97–116. Springer, Berlin, Heidelberg (2002). https://doi.org/10.1007/3-540-47804-3_8
9. Dalenius, T., Reiss, S.P.: Data-swapping: a technique for disclosure control. *J. Stat. Plann. Inference* **6**, 73–85 (1982)
10. Hundepool, A.: The ARGUS software in CENEX. In: Domingo-Ferrer, J., Franconi, L. (eds.) *PSD 2006. LNCS*, vol. 4302, pp. 334–346. Springer, Heidelberg (2006). https://doi.org/10.1007/11930242_28
11. Gouweleew, J., Kooiman, P., Willenborg, L., de Wolf, P.P.: Post randomization for statistical disclosure control: theory and implementation. *J. Official Stat.* **14**(4), 463–478 (1998)
12. Willenborg, L., de Waal, T.: *Statistical Disclosure Control in Practice*. Springer-Verlang, New York (1996). <https://doi.org/10.1007/978-1-4612-4028-0>
13. Anwar, M.N.: Microaggregation: the small aggregates method. Internal Report, Eurostat (1993)
14. Galizzi, M., Leombruni, R., Pacelli, L.: Severe work disabilities and long-lasting losses. *Labour Econ.* **85**, 102432 (2023)
15. Leombruni, R.: Interviewing Administrative Records: a conceptual map for the use of big data for economic research. *Italian J. Appl. Stat.* (2024)



A Bootstrap Approach to Account for Model Selection in Multiple Systems Estimation

Lax Chan^{1(✉)} and Bernard W. Silverman²

¹ Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Via Perrone 18, 28100 Novara, Italy

lax.chan@uniupo.it

² Department of Statistics, University of Oxford, Oxford, UK

Abstract. Multiple Systems Estimation has been used as an approach to quantify hard-to-reach or hidden populations and to inform public policy. Motivated by the application to human trafficking and to modern slavery, we review a particular methodological advancement in the frequentist regime, that makes it possible to account for model selection when finding confidence intervals for estimates.

Keywords: Bootstrap · Multiple Systems Estimation · Modern slavery

1 Introduction

It is often a difficult task to estimate the size of capture-shy and hard-to-reach populations. These situations often arise in for example, criminal justice systems and human trafficking. To obtain an estimate of the total population size is invaluable as it could raise public awareness of the severity of the issues as well as to inform public policy. Multiple systems estimation (MSE) has been introduced as a way to quantify hidden population; for an overview of the development of this discipline as well as the various application domains, one is referred to [3]. This article focuses on the application to human trafficking, and reviews material in a number of papers [4, 9–11] by the authors.

One of the standard approaches in MSE is the Poisson loglinear regression ([5]). Such an approach requires model selection, in particular on what parameters are to be included in the model, and then the estimation is conditional on the model selected for inference, see [1]. A comparison of several frequentist and Bayesian methods is illustrated in [10].

This goal of this article is to expose the latest methodological advances from a frequentist perspective, in particular the generalisation of the bootstrap approach to the well established approach to model choice using Bayes information criterion (BIC) as in [1, 2]. Since the BIC approach considers every possible model, to perform a bootstrap procedure is computationally expensive and time consuming. To overcome this bottleneck, [11] developed a modified bootstrap

approach that for each bootstrap replication, only considers models that have high-ranking BIC scores on the original data. The issue of data sparsity which is a phenomenon commonly observed in human trafficking context due to sparse or non-overlapping combination of lists is also taken into account, building on the work of [4].

The outline of the article is as follows: In Sect. 2, we will illustrate the U.K. Home Office data set on human trafficking originally analysed in [9], the model and briefly discussed the issue of data sparsity. In Sect. 3, the modified bootstrap procedure is discussed and in Sect. 4 it is applied to the U.K. Home Office dataset.

2 Data, Model and Sparsity

Table 1. LA: Local Authority, NG: Non-governmental organisation, PF: Police force: GO: Government organisations, GP: The general public

LA	X					X	X	X						X	X	X
NG		X				X			X	X	X			X	X	1 X
PF			X				X	X			X	X		X	X	X
GO				X				X	X	X		X		X	X	X
GP					X					X	X	X				
Counts	54	463	995	695	316	15	19	3	62	19	1	76	11	8	1	1

In the 2013 UK National Crime Agency strategic assessment, potential victims in modern slavery came from large number of separate source organisations, and they were crossed referenced and consolidated into five lists as shown in Table 1. Each column shows the number of cases which fall into the combination of lists indicated by the cells marked by the crosses.

Each possible combination of lists is a subset ω of $\{1, \dots, t\}$, where t is the number of lists, ω is called the capture history, the order of ω is defined to be the order of the subset and N_ω is the number of cases with the particular capture history ω . To define a model is to specify a collection of capture histories Θ , where the t capture histories of order one and the history $\emptyset$ which corresponds to the cases that do not appear on any lists are always included. The capture counts N_ω are modelled as independent random variables with

$$N_\omega \sim \text{Poisson}(\mu_\omega); \quad \log \mu_\omega = \sum_{\theta \subseteq \omega} \alpha_\theta,$$

where the sum is taken over all subsets θ of ω . The parameters can be estimated by the generalised linear model approach. We will restrict attention to hierarchical models, meaning that if $\theta \in \Theta$, then so are all subsets θ' of θ . The goal is to estimate the dark figure $N_\emptyset$ (which corresponds to the capture history $\emptyset$) and hence the total population size.

The datasets commonly observed in the context of human trafficking are often sparse in the sense that some of the capture counts N_ω are zero (e.g. the count for the combination LA:GP is 0 in Table 1) and this might lead to non-existence of maximum likelihood estimate. This was discussed in [4] as an issue that should be rigorously dealt with. Fienberg and Rinaldo [7] showed that to check the existence is equivalent to solving a linear programming problem and we refer the readers to [11] on how to economise this check.

3 BICs, Bootstrap and Models Considered

The Bayesian information criterion (BIC) [8], is one of the standard approaches to model selection and such an approach is implemented in [1] in MSE. The BIC value BIC_Θ for any particular model Θ is defined as

$$\text{BIC}_\Theta = |\Theta| \log N_{\text{total}} + 2 \sum_{\omega \notin \emptyset} (\hat{\mu}_\omega - N_\omega \log \hat{\mu}_\omega + \log N_\omega),$$

where $|\Theta|$ is the number of parameters in the model, N_{total} is the number of cases actually observed, $\hat{\mu}$ is the expected capture count estimates and the expression involved in the summation is maximum likelihood estimates. Denote $\mathcal{H}_{t,l}$ to be the set of all hierarchical models for t lists allowing parameters/interactions effects of order up to l , to apply the BIC method, the model in $\mathcal{H}_{t,l}$ with the smallest BIC_Θ value is chosen.

The bootstrap approach is a natural way to take into account model selection, because the entire estimation procedure including the choice of the model can be carried out on each of the bootstrap replications. The approach being used here is the bias-corrected and acceleration method introduced in [6]. The bootstrap replications are obtained by simulating from a multinomial distribution with N_{total} trials and event probabilities proportional to the individual N_ω . For more details on the bias correction and acceleration parameters and how to obtain the resulting confidence intervals, the readers are referred to [11].

The full bootstrap approach can be extremely computationally expensive, if we take into account all the models in the bootstrap replications. One way to mitigate this computational burden is to approximate the bootstrap using a smaller number of models. The idea is to apply fix a number n_{top} and apply the BIC condition to the original data and define $\mathcal{H}_{t,l}[n_{\text{top}}]$ to be the set of Θ in $\mathcal{H}_{t,l}$ giving the n_{top} smallest values of BIC_Θ . The key step in the approximation is then to restrict the bootstrap and jackknife choice in the bias-corrected and acceleration method considering only these n_{top} models. The intuition for such a restriction is that if the model scored extremely badly for the original data, it is unlikely that it would end up having the best BIC scores in the replication.

4 Application to the U.K. Data

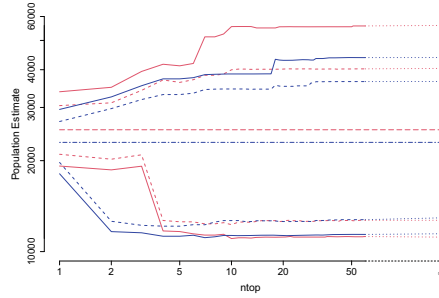


Fig. 1. Estimates and confidence interval for the U.K. 5 list data. Blue lines: maximum order 2; red lines: maximum order 4. Dot-dashed horizontal lines: estimate; short dash lines: 80% confidence interval; solid lines: 95% confidence interval. (Color figure online)

The U.K. five-list data was introduced in Table 1 in Sect. 2. This dataset was discussed in detail in [2] and subsequently, more detailed analyses were carried out in for example [10]. We applied the modified bootstrap as sketched out in Sect. 3 to this dataset, in particular to investigate the effect of varying the number of top models considered, i.e. n_{top} .

Figure 1 shows the effect of varying n_{top} , where $n_{\text{top}} = \infty$ corresponds to a full calculation taking into account all possible models of the corresponding orders (1024 models of maximum order 2 and 6893 models for maximum order 4). One can observe that for both values of maximum order, the results when $n_{\text{top}} = 60$ are practically the same. This shows that the modified bootstrap procedure introduced provide large computational gains whilst having little effect on the inference.

5 Discussion

MSE can often lead to the need of choosing from a large number of models, it is shown here that the proposed procedure has large computational gains whilst impacting little on the inference. In this work, the BIC is used and investigated but the proposed approach can be adapted to other information criteria. It would be of interest in future work to investigate the inference performance by comparing different information criteria and the method set out would make that much more feasible.

Acknowledgement. L. Chan holds a career grant supported by the European Commission - FSE REACT-EU, PON Ricerca e Innovazione 2014-2020 and is a member of Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. Baillargeon, S., Rivest, L.-P.: Recapture: loglinear models for capture-recapture in R. *J. Stat. Softw.* **19**(5), 1–31 (2007)
2. Bales, K., Hesketh, O., Silverman, B.: Modern slavery in the UK: how many victims? *Significance* **12**(3), 16–21 (2015)
3. Bird, S.M., King, R.: Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Ann. Rev. Stat. Appl.* **5**(1), 95–118 (2018)
4. Chan, L., Silverman, B.W., Vincent, K.: Multiple systems estimation for sparse capture data: inferential challenges when there are non-overlapping lists. *J. Amer. Statist. Assoc.* **116**, 1297–1306 (2021)
5. Cormack, R.M.: Log-linear models for capture-recapture. *Biometrics* **45**(2), 395–413 (1989)
6. Efron, B., Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**, 54–75 (1986)
7. Fieberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models: supplementary material (2011)
8. Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
9. Silverman, B.W.: Modern slavery: an application of Multiple Systems Estimation. UK. Published in conjunction with the UK Government Modern Slavery Strategy, Home Office (2014)
10. Silverman, B.W.: Model fitting in multiple systems analysis for the quantification of modern slavery: classical and Bayesian approaches. *J. R. Stat. Soc. Ser. A* **183**, 691–736 (2020)
11. Silverman, B.W., Chan, L., Vincent, K.: Bootstrapping multiple systems estimates to account for model selection. *Stat. Comput.* **34**, 44 (2024)



COVID-19 and Uncertainty in Italy. Exposure to High Mortality and Fertility, the Case Study on the City of Bergamo

Patrizio Lodetti^(✉) and Victoria San Juan Bernuy

Department of Statistics, Computer Science, Applications “G. Parenti”, University of Florence,
Viale Morgagni, 59, 50134 Firenze, Italy
patrizio.lodetti@unifi.it

Abstract. The COVID-19 pandemic has had a significant impact on various aspects of daily life, including reproductive behaviour. However, it is important to note that the impact of the pandemic on fertility is multifaceted and varies between and within countries and regions. This study explores the association between COVID-19 lockdowns, and exposure to death, on the birth outcomes in Italy, with a specific focus on Bergamo. Bergamo is an interesting case study as it was one of the most impacted city by the virus in the Western world. Preliminary results show that during the months of lockdown and high mortality, the number of conceptions decreases, and, conversely, during periods of reopening and stabilization of deaths the number of conceptions increases. Our results seem to corroborate the argument that the perception of uncertainty has an important impact on fertility dynamics, especially during a pandemic (and in general during catastrophic events).

Keywords: covid-19 · uncertainty · fertility

1 Theoretical Background

The COVID-19 pandemic, which emerged in late 2019 and quickly spread globally, has had profound and wide-ranging effects on various aspects of daily life and demographic patterns, including fertility dynamics. Indeed, the uncertainty brought about by the pandemic influenced childbearing plans, with implications for birth outcomes (Albeitawi et al., 2022; Bruno, Luppi, & Rosina, 2022; Guetto, Bazzani, & Vignoli, 2022; Sobotka et al., 2022; Zimmerman et al., 2022; Aassve et al., 2020). The impact of the pandemic on fertility dynamics is a complex and multi-faceted issue that varies between and within countries (Vanella et al., 2023). The academic debate has pointed to several factors that may contribute to these variations, including differences in government policies, cultural attitudes toward childbearing, and economic conditions (Brouard, Vasilopoulos, & Becher, 2020; Arpino, Luppi, & Rosina, 2021). However, one factor that has received limited attention in the literature is the role of uncertainty, which has been exacerbated by the pandemic.

The COVID-19 pandemic has indeed triggered a deep sense of uncertainty. The exposure to death and the loss of loved ones due to the virus, engendered to many emotional distress, and grief (Albeitawi et al., 2022; Aassve, Le Moglie, & Mencarini, 2021). Moreover, the implementation of lockdown measures to contain the spread of the virus, and the continuous postponement of its end, has generated feeling of helplessness. In addition, many have lost their jobs, and were exposed to financial strain and economic hardships, increasing their feelings of uncertainty about the future.

Overall, the pandemic has disrupted daily routines, social interactions, and access to healthcare. Clearly, all together, these factors have influenced reproductive decision-making both postponing fertility plans, and accelerating them because of family proximity (Bruno, Luppi, & Rosina, 2022).

Previous research has shown that uncertainty can play a significant role in shaping fertility decisions. For instance, it has been found that in situations of uncertainty or natural disasters with a sharp increase in mortality, short-term changes in fertility result in a reduction of births (Marteletto et al., 2020; Sobotka, Skirbekk, & Philipov, 2011). Studies conducted after the great recession of 2008 have confirmed this trend (Sobotka, Skirbekk, & Philipov, 2011; Wilde, Chen, & Lohmann, 2020). Concerning the specific case of COVID-19, Sobotka et al. (2022) found that as a result of the first wave of COVID-19 infections, fertility declined in December 2020 and January 2021, clearly related to a reduction in conceptions in the early months of the pandemic.

Government policies and cultural attitudes have also played a significant role in shaping fertility intentions and outcomes during the pandemic (Sobotka et al., 2022). However, also media played an important role in influencing fertility behaviors. For example, a study conducted in Jordan by Albeitawi et al. (2022) highlighted that the potential risks of pregnancy during the pandemic spread by the media was the main cause for the postponement of fertility plans. Also Guetto et al. (2021) conducted a research on the impact of media coverage on fertility. This study investigated how media portrayal of the economy affects people's decisions regarding having children in Italy. The findings revealed that economic news presented in the media has a notable influence on the likelihood of conception. Positive economic news was associated with increased fertility rates, whereas negative news decreased the probability of conceiving.

In short, COVID-19 has had important effects on demographic behaviours, not solely due to the pandemic itself but also due to individuals' perception of uncertainty about the future, which is broadly tied by the media and their narratives.

2 Italian Context and Hypotheses

The experience of the lockdowns and the high mortality are two relevant factors of uncertainty about the future (especially during the pandemic period) and are supposed to drive a decline in fertility. Bergamo, a city located in Italy, was one of the most severely impacted areas by the COVID-19 pandemic in the Western world, with high death rates and stringent lockdown measures. Therefore, it presents a unique case for studying the relationship between COVID-19, uncertainty, and birth outcomes.

Italy as a whole was subjected to stringent measures of restraint and closures (albeit with different extent and rules in its regions). Especially during lockdown periods, the

media significantly increased their reach and were sounding boards for government decisions as well as mortality trends, becoming key players in the framing of the reality and in the creation of future imaginaries.

Consequently, we defined two hypotheses. First, we expect that lockdown periods and periods of high mortality had a negative effect on conceptions. Second, we expect the negative effect of deaths on conceptions to be particularly intense in Bergamo and the provinces of the Lombardy region.

3 Data and Methods

To develop this project, we used administrative data pertaining to the monthly number of births and deaths for each of the Italian provinces.

The dependent variable is a relative measure of the number of births, covering a period that goes from December 2020 to December 2021. Specifically, for each month of this timeframe, the dependent variable has been calculated as a ratio between the number of births of the month and the mean value of that month over the previous 5 years. From now, we will name this variable 'weighted births'. The two explanatory variables, on the other hand, refer to the months of conception, covering the period from March 2020 to March 2021 (the pandemic's peak period). The first explanatory variable is a dichotomous one that asserts whether there was a lockdown situation during the month of conception or not. The second one is a relative measure of the number of deaths. For each month that goes from March 2020 to March 2021, as done previously, we calculated the ratio between the number of deaths and the mean value of that month over the previous 5 years. From now, we will name this variable 'weighted deaths'. Using deviations from the historical average allows to have comparable values between the different provinces and to take under control possible distortions due to the seasonality of these demographic phenomena.

With regard to the methodology, considering the nature of the data and the design of the analysis, we have used linear panel models with random effects.

4 Findings

In this paper, we aim to provide an innovative contribution to the literature which focuses on the relationship between uncertainty and demographic dynamics. Specifically, we do it by examining the association between COVID-19, lockdowns, and exposure to death on birth outcomes in Italy and specifically in all its provinces, giving a particular attention to Bergamo.

In Table 1, we can see the linear panel model estimating the associations between the explanatory variables and the weighed births. In the model we also included the weighed deaths variable lagged and we controlled for Provinces (not showed for a matter of space).

What can be seen is that lockdown policy is associated with a reduction of the weighted births. The same association occurs, with a similar magnitude, also when the number of weighted deaths increase in the month of conception or in the month before, supporting the first hypothesis and, more in general, the idea that that uncertainty affects negatively fertility dynamics.

Table 1. Linear panel model estimating effects on the Weighed Births

	Estimate
Lockdown	−0.0337*** (0.0057)
Weighed Deaths	−0.0227** (0.0081)
Lagged Weighed Deaths	−0.0300*** (0.0083)
Period (Year-Month)	0.0032*** (0.0009)
Constant	−2.4556*** (0.6379)
Observations	1272

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Robust St. Errors in brackets.

Model controlled by variable Province.

In Table 2, we can see the results of a model we developed to test the second hypothesis. In addition to the previous model, we included an interaction effect between provinces and the weighted mortality variable. This was done to capture a differential effect between areas most affected by the pandemic, where we expect a strong negative association, and those less affected by the pandemic, where we still expect a negative association but weaker. For each region, we show only the two provinces with the largest and smallest effects, respectively.

The first two findings are that Bergamo is not the province with the most negative effect and that there are provinces where the effect is positive. One possible interpretation is that in provinces where mortality was low and therefore the perception of uncertainty was lower, the lockdown also provided couples with the opportunity to spend more time together and consider conceiving a child, triggering mechanisms that cannot be captured by the model. Instead, in provinces where mortality was high, mortality had a direct effect on reducing births. In support of this, it should be emphasized that in Lombardy, the region most affected by the pandemic, the effect of mortality was negative and of intense magnitude in all provinces.

Following these results, we must reject the second hypothesis, even though there are elements that still capture the relationship between uncertainty, expressed in terms of lockdown and high mortality, and the decrease in births.

To improve this paper, the future steps involve developing a more sophisticated analytical model, and specifically a multilevel model with random slope and random intercept. Moreover, to broaden the goal of this study and gain a better understanding of

Table 2. .

	Estimate	Robust St. Error
Lockdown	−0.0338***	0.006
Weighed Deaths	0.0076	0.0105
Lagged Weighed Deaths	−0.0290***	0.0083
Period (Year-Month)	0.0029**	0.001
Province # Weighed Deaths	0	(.)
Vercelli # Weighed Deaths	−0.0111**	0.004
Novara # Weighed Deaths	−0.1536***	0.0029
Genova # Weighed Deaths	−0.0753***	0.0033
La Spezia # Weighed Deaths	0.1813***	0.0052
Lodi # Weighed Deaths	−0.0646***	0.0088
Bergamo # Weighed Deaths	−0.0578***	0.0074
Trento # Weighed Deaths	0.0080*	0.0036
Treviso # Weighed Deaths	0.0532***	0.0023
Venezia # Weighed Deaths	−0.0350***	0.0034
Padova # Weighed Deaths	−0.0542***	0.0049
Udine # Weighed Deaths	0.0296*	0.0129
Piacenza # Weighed Deaths	0.0160*	0.0082
Bologna # Weighed Deaths	−0.0357***	0.0047
Pesaro e Urbino # Weighed Deaths	−0.1141***	0.0051
Ascoli Piceno # Weighed Deaths	0.0075	0.0175
Pisa # Weighed Deaths	−0.0410**	0.0128
Grosseto # Weighed Deaths	0.1404***	0.0061
Perugia # Weighed Deaths	−0.0943***	0.0214
Terni # Weighed Deaths	0.0677***	0.0092
Roma # Weighed Deaths	−0.1590***	0.0156
Frosinone # Weighed Deaths	0.1213***	0.0104
Benevento # Weighed Deaths	0.1369***	0.0199
Avellino # Weighed Deaths	0.1146***	0.0153
Salerno # Weighed Deaths	−0.1556***	0.0138
Taranto # Weighed Deaths	−0.0123***	0.0065
Lecce # Weighed Deaths	0.1045***	0.0143
Potenza # Weighed Deaths	−0.0207***	0.0155

(continued)

Table 2. (continued)

	Estimate	Robust St. Error
Matera # Weighed Deaths	0.1335***	0.0079
Cosenza # Weighed Deaths	0.0725***	0.0139
Reggio di Calabria # Weighed Deaths	−0.0299***	0.0144
Enna # Weighed Deaths	0.0437***	0.0092
Ragusa # Weighed Deaths	−0.0258*	0.012
Sassari # Weighed Deaths	0.0240*	0.0104
Cagliari # Weighed Deaths	−0.2893***	0.0059
Constant	−2.2359**	0.73
Observations	1272	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Model controlled by variable Province.

the differential effect of mortality across various provinces, it will be necessary to consider the media coverage of COVID-related news, especially those concerning COVID mortality bulletins

References

Aassve, A., Cavalli, N., Mencarini, L., Plach, S., Sanders, S.: Early assessment of the relationship between the COVID-19 pandemic and births in high-income countries. *Proc. Natl. Acad. Sci.* **118**(36), e2105709118 (2021)

Aassve, A., Le Moglie, M., Mencarini, L.: Trust and fertility in uncertain times. *Popul. Stud.* **71**(1), 19–36 (2021)

Aassve, A., Cavalli, N., Mencarini, L., Plach, S., Livi Bacci, M.: The COVID-19 pandemic and human fertility. *Science* **369**(6502), 370–371 (2020). <https://doi.org/10.1126/science.abc9520>

Adsera, A., Menendez, A.: Fertility changes in Latin America in periods of economic uncertainty. *Popul. Stud.* **65**(1), 37–56 (2011)

Albeitawi, S., Al-Alami, Z., Khamaiseh, K., Al Mehaisen, L., Khamees, A., Hamadneh, J.: Conception preferences during COVID-19 pandemic lockdowns. *Behav. Sci.* **12**(5), 144 (2022). <https://doi.org/10.3390/bs12050144>

Arpino, B., Luppi, F., Rosina, A.: Changes in fertility plans during the COVID-19 pandemic in Italy: the role of occupation and income vulnerability (2021)

Brouard, S., Vasilopoulos, P., Becher, M.: Sociodemographic and Psychological Correlates of Compliance with the COVID-19 Public Health Measures in France. *Can. J. Pol. Sci./Revue Canadienne De Science Politique* **53**(2), 253–258 (2020)

Bruno, A., Luppi, F., Rosina, A.: Motivations for changing fertility plans and behaviours during the COVID-19 pandemic in Italy. *J. Family Stud.* **29**(5), 2268–2293 (2022)

Guetto, R., Bazzani, Vignoli, D.: Narratives of the future and fertility decision-making in uncertain times. An application to the COVID-19 pandemic. *Vienna Yearbook of Popul. Res.* **20**(1), 1–38 (2022)

Guetto, R., Vignoli, D., Bazzani, G.: Marriage and cohabitation under uncertain- ty: the role of narratives of the future during the COVID-19 pandemic. *Eur. Soc.* **23**(1), 674–688 (2021)

- Marteletto, L., Guedes, G., Coutinho, R., Weitzman, A.: Live births and fertility amid the Zika epidemic in Brazil. *Demography* **57**(3), 843–872 (2020)
- Nitsche, N.: Emotion and Fertility in Times of Disaster: Conceptualizing Fertility Responses to the COVID-19 Pandemic and Beyond. Max Planck Institute for Demographic Research (2022). <https://policycommons.net/artifacts/2655943/emotion-and-fertility-in-times-of-disaster/3678928/>
- Sobotka, T., et al.: From bust to boom. Birth and fertility responses to the COVID- 19 pandemic (2022)
- Sobotka, T., Skirbekk, V., Philipov, D.: Economic recession and fertility in the developed world. *Popul. Dev. Rev.* **37**(2), 267–306 (2011)
- Vanella, P., Greil, A.L., Deschermeier, P.: Fertility response to the COVID-19 pandemic in developed countries—On pre-pandemic fertility forecasts. *Comp. Popul. Stud.* **48**, 19–45 (2023)
- Vignoli, D., Guetto, R., Bazzani, G., Pirani, E., Minello, A.: A reflection on economic uncertainty and fertility in Europe: the narrative framework. *Genus* **76**, 1–27 (2020)
- Vignoli, D., Minello, A., Bazzani, G., Matera, C., Rapallini, C.: Narratives of the future affect fertility: evidence from a laboratory experiment. *Eur. J. Popul.* **38**(1), 93–124 (2022)
- Wilde, J., Chen, W., Lohmann., S.: COVID-19 and the future of US fertility: What can we learn from Google? IZA discussion paper, 13776 (2020)
- Zimmerman, L.A., et al.: Stability and change in fertility intentions in response to the COVID-19 pandemic in Kenya. *PLOS Glob. Public Health* **2**(3), 147 (2022)



Enhancing Graphical Interpretability for Functional Random Survival Forest

Giuseppe Loffredo¹✉, Elvira Romano¹, and Fabrizio Maturo²

¹ Department of Mathematics and Physics, University of Campania “Luigi Vanvitelli”, Caserta, Italy

`{giuseppe.loffredo,elvira.romano}@unicampania.it`

² Faculty of Technological and Innovation Sciences, Universitas Mercatorum, Rome, Italy

`fabrizio.maturo@unimercatorum.it`

Abstract. Functional Random Survival Forest (FRSF) has emerged as a powerful tool for analyzing functional data by starting from time-to-event data in survival analysis. However, interpreting and explaining this black-box model. This article deepens into methods for improving interpretability of Functional Survival Trees (FSTs) and explainability of FRSF. Introducing censored functional data (CFD), our approach aims to show the underlying mechanisms for survival predictions. We explore techniques for showing the significance of functional features and their contributions to survival outcomes, via the extension of the Permutation Feature Importance (PFI) tool to functional principal component scores. Moreover, we extend the concept of separation curves to survival problems with suitable interpretations. This approach aims to elucidate the model’s decision-making process and highlight key factors influencing survival predictions.

Keywords: functional data analysis · survival analysis · functional random survival forest

1 Introduction

In machine learning, non-parametric methods offer flexible and robust solutions for analyzing complex data without strict assumptions making them indispensable tools across various fields of research and practical applications. Among them, different methods are used to make predictions such as classification trees (CT), Bagging, and Random Forest (RF), where the focus is to handle discrete data points as predictors [1, 2, 6]. Recent advancements have seen the integration of a functional approach into decision tree methods, enabling the handling of functional data as features [5]. This innovation improves predictive accuracy and deepens information about dynamic datasets, promising significant improvements in various fields of data analysis and machine learning. In survival analysis, the concept of Functional Random Survival Forests (FRSF) has emerged

as a promising approach extending the framework of the Random Survival Forest (RSF) to accommodate the analysis of Censored Functional Data (CFD) as features in the predictive model [4]. However, when dealing with survival analysis, considering the CFD, the principal problems are to interpret single FST and explain FRSF. To address these problems, our study focuses on proposing instruments and methods for the interpretability of individual trees and explainability of the ensemble. This involves defining extending the concept of separation curves in the survival analysis context and examining results at each node to understand how the model partitions the feature space in terms of survival probability.

2 Random Survival Forest for Functional Data

Functional Random Survival Forest (FRSF) is a machine learning algorithm used in survival analysis. It's an extension of Random Survival Forest (RSF) with functional features for predicting time-to-event outcomes. Differently from classic RSF, to overcome the curse of dimensionality, data are reconstructed by using functional principal component decomposition (FPC) and the principal scores are considered as new features in the model. The FRSF can be explained in a few steps: at the beginning, the algorithm creates B bootstrap samples, with replacement, from the starting dataset. For each bootstrap sample, it builds a survival tree choosing in each node a set of scores as feature candidates for splitting. Specific survival criteria are used for time-to-event outcomes such as Log-rank, Log-rank score, and Bs-gradient. After multiple survival trees are built, the predictions are made by aggregating the predictions from individual trees. In survival analysis, the predicted outcome is typically the survival probability or the expected survival time. However, there are still some aspects that can be explored and developed such as the improvement of accuracy of functional classifiers, and the introduction of understandable graphical tools for the interpretation of classification rules.

2.1 Functional Data Analysis for Survival Time Data

The aim of the Functional Data Analysis (FDA) is to handle data functions as single objects. In practical applications, however, functional data are often observed as a series of point data, where a function $f(x)$ is defined by starting to discrete observations denoted as (x_j, t_j) , where $x_j \in \mathbb{R}$ and t_j represents the value of the function computed at the point x_j for $j = 1, 2, \dots, T$. Generalizing this framework, we can define a set of N random function variables assuming values in a functional space L^2 , given by $X_1(t), \dots, X_N(t)$. Focusing on a Hilbert space $(L^2(\mathcal{T}), \|\cdot\|_2)$ where $\mathcal{T}$ is a time domain and $\|\cdot\|_2$ is the metric associated into functional space, we can define the distance between two functions as follows:

$$\|X_1(t) - X_2(t)\|_2 = \left(\int_{\mathcal{T}} |X_1(t) - X_2(t)|^2 dt \right)^{\frac{1}{2}}. \quad (1)$$

In FDA the first step is to convert observed values x_{ij} for each unit $i = 1, 2, \dots, N$ to a functional form using a finite representation in a fixed basis system, given by:

$$X_i(t) = \sum_{k=1}^K c_{ik} \phi_k(t) \quad (2)$$

where $\mathbf{c}_i = (c_{i1}, \dots, c_{iK})^T$ is the vector of coefficients, and $\phi_k(t)$ is the k -th basis function from a subset of K functions used to approximate the full basis expansion. Other approaches can be used to build functional variables such as the Functional Principal Components (FPCs) decomposition, where functional data can be approximated as:

$$X_i(t) = \sum_{z=1}^Z \nu_{iz} \xi_z(t), \quad (3)$$

where K is the total number of FPCs, $\xi_z(t)$ is the eigenfunction for the generic function X_i , and ν_{iz} is the principal component score. In survival analysis, the data for $i = 1, \dots, N$ subjects, during the period $\mathcal{T}$, are recorded in J_i different values, where the times T_i which denote the survival time and the times C_i which denote the censored time that it represents an event not occurred by a study time $\mathcal{T}$. The survival time data denoted by $(\mathbf{Y}_i, T_i^*, \delta_i)$ where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ_i})^T$ refers to the observed values vector for subject i with $Y_{ij} = Y(t_{ij})$, $t_{ij} \in \mathcal{T}$ for $j = 1, \dots, J_i$ observations and δ_i is an indicator function with $\delta_i = 0$ for $T_i^* = T_i$ and 0 otherwise. From these data, we introduce the Censored Functional Data (CFD) defined in [4] such to give a graphical interpretation within the tree in the FRSF. The CFD is defined as follows:

$$\begin{aligned} \mathcal{X}_i &= \{X_i(t) \in L_2(\mathcal{T}) : t \in [0, T_i^*]\} \quad \text{with} \quad T_i^* = \min(T_i, C_i) \\ \text{and} \quad \delta_i &= 1_{[T_i^* = T_i]} \quad \text{for} \quad i = 1, \dots, N. \end{aligned} \quad (4)$$

From (4) in order to obtain the FPCs we apply a fully non-parametric approach denoted as Principal Analysis Conditional Expectation (PACE) [7].

3 Graphical Interpretability for FST and Explainability for FRSF

Graphical interpretability for Functional Survival Tree (FST) and explainability for FRSF can be achieved through different tools. In the first case, our aim will be to explore the possibility of introducing a graphic interpretation by considering the survival curves defined for different groups in each ST. Starting from the concept of separation curve in the context of survival analysis we interpret the meaning of each node splitting in the single ST. The separation curve in a functional space is defined by extending the Functional Predicted Separation

Prototype (FPSP) and the Functional Empirical Separation Prototype (FESP) [5] expressed as follows:

$$FPSP_h(t) = \sum_{k \in \Omega} \nu_{0k} \xi_k(t) \quad (5)$$

$$FESP_h(t) = \min_i d_2(X_i(t), FPSP_z(t)) \quad (6)$$

where for the (5) equation ν_0 is the feature selected in node h , Ω is a set of the FPCs $\xi_k(t)$ selected in the classification rule path until the split of the h -th node. In contrast, for the (6) we have d_2 is a metric defined in L_2 space and $X_i(t)$ are the approximated curves defined by FPCA decomposition. These graphical instruments provide a means to interpret the single FST for understanding the model's behavior, identifying important survival curves in each daughter node generated from the splitting methods (log-rank test, log-rank score and Bs-gradient). In particular, we propose to plot survival curves (Keple-Meier, Cox regression and so on) and we'll analyze the CFD defined in (4) in order to evaluate the impact of predictor variables on survival outcomes and to study the shape of variation in each terminal node. By analyzing these curves and functions, we aim to classify individuals into different risk groups or categories based on their predicted outcomes. Once that the FRSF model is constructed, the next step involves assessing feature importance for explainability. We use the Permutation Feature Importance (PFI), which measures the increase in prediction error after randomly permuting feature values, thus disrupting the feature-outcome relationship. This method allows us to gauge the contribution of each feature to the model's predictive performances in terms of C-index. The method is based on [3] where in the functional context, we consider the feature-score matrix V_{perm} , where the columns are the FPCs calculated on CFD by FPACE decomposition. The FPCs are individually permuted to address their importance. The process is repeated many times for each FPCs to obtain a distribution of increase/decrease of model performance.

References

1. Breiman, L.: Bagging predictors. *Mach. Learn.* (1996). <https://doi.org/10.1007/bf00058655>
2. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2004)
3. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously (2018). <http://arxiv.org/abs/1801.01489>
4. Loffredo, G., Romano, E., Maturo, F.: Random survival forest for censored functional data. In: SDS-Conference, Palermo, Italy (2024)
5. Maturo, F., Verde, R.: Supervised classification of curves via a combined use of functional data analysis and tree-based methods. *Comput. Stat.* **38**, 419–459 (2023)
6. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* (1986). <https://doi.org/10.1023/A:1022643204877>
7. Yao, F., Müller, H.-G., Wang, J.-L.: Functional linear regression analysis for longitudinal data, 2873–2903 (2005)



Uncertainty in Sampling Designs for Non-probability Samples

Pier Luigi Conti¹ and Daniela Marella²(✉)

¹ Dipartimento di Scienze Statistiche, Sapienza Università di Roma,
Piazzale Aldo Moro 5, 00185 Rome, Italy

pierluigi.conti@uniroma1.it

² Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma,
Piazzale Aldo Moro 5, 00185 Rome, Italy

daniela.marella@uniroma1.it

Abstract. Non-probability samples involve some form of arbitrary selection of units into the sample, and, as a matter of fact, inclusion probabilities are unknown. Hence, it is not possible to apply probability randomization theory to make inference about the finite population parameters. In this paper the concept of uncertainty on data generating model, resulting from the lack of knowledge of the sampling design acting in the non-probability sample is discussed. A measure of uncertainty is introduced and its asymptotic proprieties are evaluated.

Keywords: Nonignorable designs · selection bias · uncertainty

1 Introduction

In the last years the relevance of non-probability sampling in surveys is considerably increased because of the availability of massive datasets (Big data), whose use is probably the most challenging problem National Statistical Offices have to face.

In probability sampling every unit in a finite population has a known non-zero probability of being selected. Non-probability sampling involves some form of arbitrary selection of units into the sample, inclusion probabilities are unknown and then it is not possible to apply probability randomisation theory to make inference about the finite population parameters. Furthermore, the unknown selection process is frequently selective with respect to the target population, so that estimates of population characteristics may be subject to serious selection bias [2].

Suppose that we have two samples A and B , where B is a non-probability sample and A is an independent probability sample. Let $\mathcal{Y}$ be the study variable and $\mathcal{X} = (\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p)^T$ a vector of p auxiliary variables; $(\mathcal{X}, \mathcal{Y})$ are available from B , while $\mathcal{X}$ is available from survey data A . When a probability sample A is available, two approaches can be considered. The first one consists in the propensity score adjustment [7], where the probability of being included in the

non-probability sample B is estimated from sample A through the covariates $\mathcal{X}$. In the second approach models are fitted to B to predict the variable of interest $\mathcal{Y}$ for the units in the probability sample A (mass imputation), see [8]. Both the aforementioned approaches assume that the selection mechanism for the non-probability sample B is ignorable after controlling on covariates $\mathcal{X}$. Formally, the sampling mechanism for the non-probability sample B is noninformative (and hence ignorable) after controlling on $\mathbf{x}$ if

$$p(\delta_i = 1 | \mathbf{x}_i, y_i) = p(\delta_i = 1 | \mathbf{x}_i) \quad (1)$$

for each $\mathbf{x}_i$. If the sampling mechanism for sample B is not ignorable, the inclusion probabilities depend on the values of the variable $\mathcal{Y}$ even after conditioning on the covariates $\mathcal{X}$, so that the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data does not coincide with the model holding for the population. See [5,6] and references therein for discussion of the notion of informative sampling.

In order to account for a nonignorable selection mechanism, as in [3,4], we assume that the sample membership indicators of B can be observed in the probability sample A .

In this paper the concept of uncertainty on data generating model, resulting from the lack of knowledge of the sampling design acting in the non-probability sample, is discussed. Such an uncertainty can be reduced due to the availability of extra-sample information.

The paper is organized as follows. In Sect. 2, the class of plausible distributions for the variable of interest is defined on the basis of the sample data available in the non-probability sample B and in the reference survey A . Next, an uncertainty measure quantifying the size of the aforementioned class is introduced, an estimator is defined and its asymptotic proprieties are analyzed.

The class of plausible distributions can be reduced if extra-sample information on the unknown informative sampling design in B , as well as on the distribution of the variable of interest $\mathcal{Y}$, is available, as briefly discussed in Sect. 3.

2 Uncertainty in Sampling Designs

Let $\mathcal{U} = \{u_1, \dots, u_N\}$ be a finite population of size N . Let $\mathcal{Y}$ be the variable of interest and $\mathcal{X} = (\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p)^T$ a set of p auxiliary variables. We assume that both $\mathcal{Y}$ and $\mathcal{X}$ are categorical variables, with J and I categories, respectively. The symbols $y_i, \mathbf{x}_i$ are used to denote the values of $\mathcal{Y}$ and $\mathcal{X}$ for the population unit $u_i, i = 1, \dots, N$. The joint, marginal and conditional *population probability mass function* (*ppmf*, for short) are denoted by

$$p_N(y, \mathbf{x}), \quad p_N(y), \quad p_N(y|\mathbf{x}) = \frac{p_N(y, \mathbf{x})}{p_N(\mathbf{x})}. \quad (2)$$

In addition, a superpopulation model is adopted, with $(y_i, \mathbf{x}_i)$ s realizations of independent and identically distributed (*i.i.d.*) random variables (r.v.s) $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, N$. The symbols

$$p(y, \mathbf{x}), p(y), p(\mathbf{x}), p(y|\mathbf{x}) = \frac{p(y, \mathbf{x})}{p(\mathbf{x})} \quad (3)$$

denote the corresponding joint, marginal, and conditional probability mass functions (*pmfs*, for short), respectively. Next, let A, B be two independent samples of size n_A, n_B , respectively. Sample A is a probability sample where only variables $\mathbf{X}$ are observed, whilst sample B is a non-probability sample where both $\mathcal{Y}$ and $\mathbf{X}$ are observed. Let D_i, δ_i be the sample membership indicators of unit u_i , in samples A and B , respectively, namely:

$$D_i = \begin{cases} 1 & \text{if } i \in A \\ 0 & \text{otherwise} \end{cases}, \quad \delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \dots, N. \quad (4)$$

Furthermore, the sample A is selected according to a known, probabilistic, non-informative design, with first order inclusion probabilities π_i^A .

We assume that the sample indicators δ_i s are observed in sample A . Such an assumption is more plausible in the Big data context, since the non-probability sample B is large, and, with high probability, the probability sample A does have a non empty intersection with B .

Then, the values $(\delta_i, \mathbf{x}_i)$ are observed in sample A and the values $(y_i, \mathbf{x}_i)$ are observed in the non-probability sample B . Roughly speaking, B contains information on $(\mathbf{X}, \mathcal{Y})$, but the sampling mechanism is unknown. In sample A the study variable of interest $\mathcal{Y}$ is not observed, but information on δ is available.

The estimation of the joint *pmf* $p(\mathbf{x}, y)$ requires the estimation of both the marginal *pmf* $p(\mathbf{x})$, and the conditional *pmf* $p(y|\mathbf{x})$. The former can be estimated from the probability sample A , whilst the latter can be estimated from the non-probability sample B . However, since the sampling design acting in B is unknown and possibly informative, estimates obtained under assumption (1) could be affected by serious bias. By the law of total probability, the *pmf* $p(y|\mathbf{x})$ can be written as

$$p(y|\mathbf{x}) = p(\delta = 1|\mathbf{x})p(y|\mathbf{x}, \delta = 1) + p(\delta = 0|\mathbf{x})p(y|\mathbf{x}, \delta = 0), \quad (5)$$

where $p(y|\mathbf{x}, \delta = 1)$ and $p(y|\mathbf{x}, \delta = 0)$ are the *sample* and the *sample complement* distribution, respectively. Note that, $p(y|\mathbf{x}), p(y|\mathbf{x}, \delta = 1), p(y|\mathbf{x}, \delta = 0)$ do coincide if and only if $p(\delta = 1|\mathbf{x}, y) = p(\delta = 1|\mathbf{x})$, namely if the sampling design is ignorable. If $p(\delta = 1|\mathbf{x}), p(y|\mathbf{x}, \delta = 1)$ were known, the only valid statement about $p(y|\mathbf{x})$ would be in the identification region

$$H[p(y|\mathbf{x})] = \{p(\delta = 1|\mathbf{x})p(y|\mathbf{x}, \delta = 1) + p(\delta = 0|\mathbf{x})\gamma, \gamma \in \Gamma_{\mathbf{x}y}\} \quad (6)$$

where $\gamma = p(y|\mathbf{x}, \delta = 0)$, and $\Gamma_{\mathbf{x}y}$ denotes the space of all sample complement models. The class (6) identifies the set of all plausible distributions for $p(y|\mathbf{x})$

compatible with our knowledge of $p(\delta = 1|\mathbf{x})$ and $p(y|\mathbf{x}, \delta = 1)$. Notice that γ depends on $p(\delta = 0|\mathbf{x}, y) = 1 - p(\delta = 1|\mathbf{x}, y)$. Hence, Γ_{xy} can be re-written as

$$\Gamma_{xy} = \left\{ p(\delta = 1|\mathbf{x}, y) : p(\delta = 1|\mathbf{x}) = \sum_y p(\delta = 1|\mathbf{x}, y)p(y|\mathbf{x}) \right\}. \quad (7)$$

It can be interpreted as the class of all possible sampling designs that could have generated the non-probability sample B .

When the ignorability assumption (1) is satisfied, the class (6) collapses into a single *pmf*, namely $p(y|\mathbf{x}) = p(y|\mathbf{x}, \delta = 1)$. In all other cases the sampling design generating B cannot be identified, and the same holds for the *pmf* $p(y|\mathbf{x})$.

An uncertainty measure on $p(y|\mathbf{x})$ (*conditional uncertainty measure*) can be defined as

$$U^x(p(y|\mathbf{x})) = \frac{1}{J} \sum_{j=1}^J (1 - p(\delta = 1|\mathbf{x})) = 1 - p(\delta = 1|\mathbf{x}). \quad (8)$$

Finally, from (8), uncertainty on the marginal probability $p(y)$ can be defined as a weighted mean of the conditional uncertainty measures (8)

$$U(p(y)) = \sum_{\mathbf{x}} p(\mathbf{x})(1 - p(\delta = 1|\mathbf{x})) = 1 - p(\delta = 1), \quad (9)$$

where $p(\mathbf{x})$ is the *pmf* of $\mathcal{X}$. Both $U^x(p(y|\mathbf{x}))$ and $U(p(y))$ take values in $[0, 1]$. With regard to $U^x(p(y|\mathbf{x}))$, it takes the value 0 if $p(\delta = 1|\mathbf{x}) = 1$, that is if, conditionally on $\mathbf{x}$, all units in the population are sampled. Analogously, it takes the value 1 if $p(\delta = 1|\mathbf{x}) = 0$, that is no sample data are available and the uncertainty is maximum. The same consideration holds for $U(p(y))$.

2.1 Estimators of the Uncertainty Measures

In this section estimators of the conditional and unconditional uncertainty measures (8), (9) are defined and their asymptotic proprieties are analyzed. First of all, on the basis of sample A the following quantities can be estimated:

- Inclusion probability $p(\delta = 1|\mathbf{x})$ by the Hájek estimator $\hat{p}(\delta = 1|\mathbf{x})$ defined as

$$\hat{p}(\delta = 1|\mathbf{x}) = \frac{\sum_{i=1}^N \frac{1}{\pi_i^A} I(\mathbf{X}_i = \mathbf{x}) \delta_i D_i}{\sum_{i=1}^N \frac{1}{\pi_i^A} I(\mathbf{X}_i = \mathbf{x}) D_i}. \quad (10)$$

- Marginal probability $p(\delta = 1)$ by Hájek estimators $\hat{p}(\delta = 1)$ defined as,

$$\hat{p}(\delta = 1) = \frac{\sum_{i=1}^N \frac{1}{\pi_i^A} \delta_i}{\sum_{i=1}^N \frac{1}{\pi_i^A}}. \quad (11)$$

As an estimator of the conditional uncertainty measure (8) it is intuitive to consider

$$\widehat{U}^x(p(y|\mathbf{x})) = 1 - \widehat{p}(\delta = 1|\mathbf{x}). \quad (12)$$

The properties of $\widehat{U}^x(p(y|\mathbf{x}))$ are analyzed in [1]. Formally,

$$\widehat{U}^x(p(y|\mathbf{x})) \xrightarrow{p} U^x(p(y|\mathbf{x})) \text{ as } n_A, N \rightarrow \infty; \quad (13)$$

$$\sqrt{N}(\widehat{U}^x(p(y|\mathbf{x})) - U^x(p(y|\mathbf{x}))) \xrightarrow{d} N(0, \sigma_H^2) \text{ as } n_A, N \rightarrow \infty. \quad (14)$$

In a similar way, as an estimator of the unconditionally uncertainty measure (9) we consider

$$\widehat{U}(p(y)) = \sum_{\mathbf{x}} \widehat{p}(\mathbf{x})(1 - \widehat{p}(\delta = 1|\mathbf{x})) = 1 - \widehat{p}(\delta = 1), \quad (15)$$

where the weights $p(\mathbf{x})$ can be estimated from the probability sample A through the Hájek estimator

$$\widehat{p}(\mathbf{x}) = \frac{\sum_{i=1}^N \frac{1}{\pi_i^A} I(\mathbf{X}_i = \mathbf{x})}{\sum_{i=1}^N \frac{1}{\pi_i^A}}. \quad (16)$$

Large sample properties of (15) and (11) are established in Proposition 1 and proved in [1].

Proposition 1. *As $N, n_A \rightarrow \infty$, $\widehat{U}(p(y)) \xrightarrow{p} U(p(y)) = 1 - p(\delta = 1)$, and $\sqrt{N}(\widehat{p}(\delta = 1) - p(\delta = 1))$ tends in distribution to a Normal variate with null mean and variance $\sigma_{H\delta}^2$.*

3 Reducing Uncertainty by Auxiliary Information

When extra-sample information is available some *pmfs* in the class (6) are no longer admissible and must be excluded, with a consequent reduction of uncertainty. In [1], two kinds of extra-sample information are considered.

- (i) Auxiliary information on the informative sampling design acting in B .
- (ii) Auxiliary information on the conditional *pmf* $p(y|\mathbf{x})$.

Notice that, information (i) and (ii) interact each other. Extra-sample information on the characteristics of sampling design acting in B reduces the class of plausible sampling design (7), as well as the class of plausible distributions $H(p(y|\mathbf{x}))$, and then the uncertainty on the data generating model. The same holds if extra-sample information regarding the conditional *pmf* $p(y|\mathbf{x})$ is available.

More specifically, in [1] constrained uncertainty measures under the above scenarios (i) and (ii) are defined and their asymptotic proprieties are analyzed. Finally, a plausible estimate for the distribution of the variable of interest is constructed and its accuracy is evaluated by the aforementioned uncertainty measure.

References

1. Conti, P.L, Marella, D.: An uncertainty based approach for dealing with selection bias in non-probability samples (2024, submitted)
2. Elliott, M., Valliant, R.: Inference for non-probability samples. *Stat. Sci.* **32**(2), 249–264 (2017). <https://doi.org/10.1214/16-STS598>
3. Kim, J.K., Wang, Z.: Sampling techniques for big data analysis in finite population inference. *Int. Stat. Rev.* **87**, S1:77–S1:91 (2019). <https://doi.org/10.1111/insr.12290>
4. Marella, D.: Adjusting for selection bias in nonprobability samples by empirical likelihood approach. *J. Official Stat.* **39**(2), 151–172 (2023)
5. Pfeffermann, D., Sverchkov, M.: Inference under informative sampling. In: Pfeffermann, D., Rao, C.R. (eds.) *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*, pp. 455–487. North Holland (2009)
6. Pfeffermann, D.: Modelling of complex survey data: why model? Why is it a problem? How can we approach it? *Surv. Pract.* **37**(2), 115–136 (2011)
7. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55 (1983). <https://doi.org/10.1093/biomet/70.1.41>
8. Yang, S., Kim, J.K., Hwang, Y.: Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Surv. Pract.* **47**(1), 29–58 (2021)



Model-Based Clustering of Nested Lifetime Data: Profiling COVID-19 Heart Failure Patients

Chiara Masci¹(✉), Andrea Cappozzo², Francesca Ieva^{1,3}, Olivia Leoni⁴,
Marco Forlani⁵, Barbara Antonelli⁶, and Anna Maria Paganoni¹

¹ MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

{chiara.masci, francesca.ieva, anna.paganoni}@polimi.it

² Department of Economics, Management and Quantitative Methods, via Conservatorio 7, 20122 Milan, Italy
andrea.cappozzo@unimi.it

³ Health Data Science Center, Human Technopole, Milan, Italy

⁴ U.O. Osservatorio Epidemiologico, DG Welfare, Regione Lombardia, Piazza Città di Lombardia 1, 20124 Milan, Italy

⁵ Regione Lombardia, Informatica SPA, Piazza Città di Lombardia n.1, 20124 Milan, Italy

⁶ Regione Lombardia, Divisione Servizi per il Welfare Regionale, Piazza Città di Lombardia n.1, 20124 Milan, Italy

Abstract. This study presents a novel multilevel cluster weighted model for lifetime data with random covariates and frailties. In the framework of model-based clustering, we extend cluster-weighted model to handle time-to-event responses and data with hierarchical structures. The proposed methodology allows to identify latent clusters of observations characterised by random covariates with different distributions, whose time-to-event dynamic and heterogeneity at the grouping level also differ. The objective function that defines the model is maximized using a stochastic EM algorithm tailored to right-censored lifetime data. The development of this method is motivated by a study on the survival of COVID-19 heart failure patients, admitted to multiple hospitals in Lombardy region. By applying the proposed method, we identify latent clusters of patients that differ in terms of clinical characteristics and, for each cluster, we investigate the survival pattern, the association of particular respiratory diseases with their death hazard and the supplementary effect of the facility.

Keywords: Cluster-weighted models · Survival data · Frailty models · EM algorithm · Health analytics

1 Introduction

The framework of survival analysis is well established and has been widely used when dealing with clinical data. In this context, the most relevant information

is the time until the event of interest occurs (time-to-event). Many books treat this matter, see e.g. [4]. The Cox model extends the traditional methodology to account for covariate contribution in a regression fashion. Here, together with the assumption of proportional hazard, observations are assumed to be independent. In particular real data contexts, this assumption does not hold and efforts have been undertaken to mitigate or relax this hypothesis. The frailty model, introduced in [7], acts on the hazard function estimating a multiplicative quantity called frailty, accounting for an additional level of hierarchy present when data are nested. This multilevel approach proves highly efficient in examining and addressing time-to-event data while accommodating for the heterogeneity in a finite number of homogeneous classes for a known hierarchy. However, in practical applications, there is often a supplementary source of heterogeneity, at the lowest level of the hierarchy (e.g., patients), that would require to incorporate covariates or frailty effects that are distinct for various patient clusters, instead of relying on a singular estimate for the entire population. Within this framework, we propose an approach that identifies clusters of patients with diverse characteristics, within which the frailty survival model dynamic is allowed to vary. To this end, we rely on multilevel Cluster Weighted Models (CWM) [2] that implement a clustering on the observations based on the marginal distributions of the covariates inside each cluster and the conditional distribution of the response given the covariates. In particular, we extend the CWM approach for hierarchical data to a time-to-event outcome. To the best of our knowledge, this work is the first that takes into account both the clustering of multilevel data and the intrinsic heterogeneity in a survival context. Previous works have been done with continuous and dichotomous responses in [1, 2]. The innovative aspect of our work resides in the possibility of a deep and more informed analysis of the hazard risk stratification, disentangling the contributes that arise from different data levels. The method is applied to predict the time-to death of Heart Failure (HF) patients affected by COVID-19 in Lombardia Region. The customization acts on the clinical characteristics of the patient and most importantly on the hospital level, offering the possibility of a scenario analysis for the medical facilities through the frailty term. The baseline and the frailty distributions are estimated in a parametric framework. The estimation of the model is carried out by means of an Expectation-Maximization algorithm tailored to right-censored lifetime data [3]. We implemented this algorithm in the R programming language [6], relying on the existing R package `parfm` [5].

2 Method

Consider a survival framework for two-levels data, i.e., patients within hospitals. Each statistical unit is identified by the triplet $(y_{ij}, \delta_{ij}, \mathbf{x}_{ij})$, where

- $j = 1, \dots, J$ is the hospital index (J hospitals),
- $i = 1, \dots, n_j$ is the individual index (n_j patients in hospital j),
- $N = \sum_{j=1}^J n_j$ is the total size of the sample,

- $t_{ij} = \min(y_{ij}, c_{ij})$, with y_{ij} and c_{ij} being the survival and censoring time,
- $\delta_{ij} = 1(t_{ij} < c_{ij})$ is the event indicator,
- $\mathbf{x}_{ij} = (\mathbf{u}_{ij}, \mathbf{v}_{ij})$ denotes the vector of covariates with $\mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ indicating the subset of continuous and categorical predictors for the ij -th unit.

It can be proven that the log-likelihood of the CWM model for nested lifetime data reads as in Eq. (1), where $\boldsymbol{\beta}$ are the covariate coefficients, $\boldsymbol{\theta}$ is the unknown parameter for the frailty distribution, h_0 and H_0 are the baseline and cumulative hazard respectively, $\mathcal{L}^{(d)}(\cdot)$ is the d -th derivative of the Laplace transform of the frailty, $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\xi(\cdot; \boldsymbol{\Pi})$ indicate the density of a Normal and a Multinomial distribution and R_{jg} contains the indexes of the observations assigned to cluster g belonging to hospital j .

$$\begin{aligned} \ell \left(\{\tau_g, \boldsymbol{\beta}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\Pi}_g, \theta_g\}_{g=1}^G \right) = & \sum_{g=1}^G \left\{ \sum_{j=1}^J \sum_{i \in R_{jg}} \log \tau_g \right. \\ & + \sum_{j=1}^J \left[\sum_{i \in R_{jg}} \delta_{ij} (\log h_0(y_{ij}) + \mathbf{x}'_{ij} \boldsymbol{\beta}_g) \right. \\ & + \log \left[(-1)^{d_{jg}} \mathcal{L}^{(d_{jg})} \left(\sum_{i \in R_{jg}} H_0(y_{ij}) \exp(\mathbf{x}'_{ij} \boldsymbol{\beta}_g); \theta_g \right) \right] \left. \right] \\ & \left. + \sum_{j=1}^J \sum_{i \in R_{jg}} \log \phi(\mathbf{u}_{ij}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \sum_{j=1}^J \sum_{i \in R_{jg}} \log \xi(\mathbf{v}_{ij}; \boldsymbol{\Pi}_g) \right\} \end{aligned} \quad (1)$$

To get this expression, we rely on the work proved in [5], adding also the effect of covariates, that are modelled accordingly to their nature. In particular, when dealing with continuous variables within cluster g , we make use of a multivariate Gaussian distribution, parameterized by the multivariate mean $\boldsymbol{\mu}_g$ of size p and its covariance matrix $\boldsymbol{\Sigma}_g$ of size $p \times p$, where p is the number of continuous covariates. Instead, we model categorical variables with a multinomial distribution. The associated parameter is denoted with $\boldsymbol{\Pi}_g$, whose components we call $\boldsymbol{\pi}_{g1}, \dots, \boldsymbol{\pi}_{gq}$, where q is the number of categorical covariates. The maximization of the log-likelihood is carried out via EM algorithm. Each term of the log-likelihood can be maximized separately. To conduct model selection we make use of the BIC criterion, that takes into account the fitting of the model penalizing its complexity.

3 Application to Lombardy Region Database

We apply the proposed methodology to the administrative database of the Lombardy Region health system. Units of analysis and groups are represented by patients and hospitals of admittance. Specifically, our goals are:

- Identify relevant latent clusters of patients using their clinical records, in order to create meaningful profiles to be studied in terms of survival.
- Investigate the effect of the structures (medical facilities) on the hazard of death of individuals with different profiles, discovering the heterogeneity given to this level of the data hierarchy.
- Analyze the effect of respiratory pathologies on the hazard of death of different profiles of patients, to understand if their survival in this context is compromised by the presence of some illness.
- Aggregate all sources of information to produce custom survival curves, with a risk stratification perspective that takes into account categories of patients, respiratory diseases and hospital of admittance.

Data Presentation and Model Definition. We focus on a cohort of patients with heart failure who have been hospitalized for COVID-19, on a time span from August 2020 to May 2023. Our dataset includes 3086 patients admitted in $J = 32$ hospitals in the Lombardy region. We consider a time window of 90 days for studying the survival time. For each patient, we select two sets of variables: the first one is represented by Age, Gender and a Modified Multisource-Comorbidities Score (MCS). The MCS is a numerical index that summarizes the clinical status of the patient, taking into account a wide range of pathologies. It has been modified by removing the respiratory diseases we are interested to study separately. These three variables are modelled according to their nature and they contribute only in defining the clusters. The second set of variables collects the presence of the three respiratory diseases we removed from the MCS: Chronic Obstructive Pulmonary Disease (COPD), Pneumonia (PNA), and Respiratory Failure (RF). They are included as binary indicator (ill/healthy) only in the regression section, to evaluate the effect on the survivability of the different categories of patients. The response variable, i.e., the time-to-death, is expressed in fraction of years.

Results. We fit the model for $G \in \{2, 3, 4, 5\}$, and two choices for the baseline, the Weibull and the Lognormal distribution. The frailty is modelled with a Gamma distribution of mean 1 and unknown variance θ_g for each cluster.

The best result in terms of BIC is obtained when $G = 3$ and a Lognormal baseline is employed. The complete set of estimated parameters is reported in Table 1. Cluster 1 contains relatively young patients with a clinical situation which is not compromised (11% of the total observations).

In left panel of Fig. 1 we represent the baseline survival function specific to each cluster, that correspond to the estimated survival function for a patient not affected by the respiratory diseases that we consider as covariates. Here, to quantify the effect of the frailty, we produce shaded bands of amplitude $\pm 1 \text{ st dev} = \sqrt{\widehat{\theta}}$, when significantly different from 0. There is a significant difference in terms of survival for the three clusters in the baseline estimates, which is coherent with the interpretation we gave to the clusters. The survival curve for each specific patient can be obtained by adding the contribution of respiratory illnesses and of the hospital to the baseline.

Table 1. Parameter estimates for the three clusters. $\hat{\mu}_{g,1}$ refers to the Age, $\hat{\mu}_{g,2}$ to the MCS, $\hat{\pi}_g$ to the Gender. θ_g is the variance of the frailty. η_g and δ_g are the characteristic parameters of the Lognormal (η_g, δ_g) baseline. Covariates order corresponding to $\hat{\beta}_g$ is COPD, PNA and RF. Asterisc indicates the parameters which are significantly different from 0 at level 5%.

g	$\hat{\mu}_g$	$\hat{\Sigma}_g$	$\hat{\pi}_g$	$\hat{\theta}_g$	$\hat{\eta}_g$	$\hat{\delta}_g$	$\hat{\beta}_g$
1	(66.02, 10.02)	$\begin{pmatrix} 10.04 & 1.00 \\ 1.00 & 56.58 \end{pmatrix}$	(0.26, 0.74)	0.67*	1.49	2.99	$\{-0.32, -0.026, 1.05^*\}$
2	(83.40, 10.52)	$\begin{pmatrix} 42.79 & -1.72 \\ -1.72 & 37.71 \end{pmatrix}$	(0.47, 0.53)	0.15*	-1.77	2.01	$\{-0.010, 0.19^*, -0.077\}$
3	(79.24, 32.17)	$\begin{pmatrix} 52.01 & -6.47 \\ -6.47 & 32.90 \end{pmatrix}$	(0.38, 0.62)	0.27	-2.66	1.45	$\{-1.03^*, 0.34, 0.85^*\}$

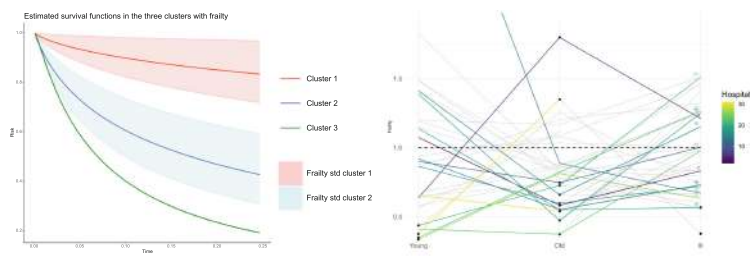


Fig. 1. Left panel: Estimated survival curves in the 3 clusters. The red line corresponds to the baseline of cluster 1, the blue curve to cluster 2 and the green curve to cluster 3. The shaded bands represent the frailty effect, when significant. Right panel: Frailties estimates in the three clusters (young, old and ill patients). Each line corresponds to one hospital. Black points identify the frailties statistically different from 1. Colored lines refer to hospitals with at least one significant frailty.

The first cluster shows the best curve in terms of survivability. Here, the presence of RF is the only significant covariate: the estimated $\hat{\beta}_{1,RF}$ is positive, therefore the risk is higher than the one of a patient without this illness. For the second cluster, the magnitude of the risk is higher than cluster 1, confirming our expectations for older patients whose life is more at risk. The presence of PNA is the only covariate with statistical significance. A patient with this pathology is more at risk. The third cluster, containing frail patients, has the highest associated risk. The covariates coefficients that are of interest in the survival are COPD and RF. While this second one has positive sign with a straightforward meaning, the $\hat{\beta}_{3,COPD}$ is negative, causing a protective effect. Probably the reason is that while RF is a complication of the clinical picture that can occur with COVID-19, COPD is a pre-existing pathology. These patients are particularly attentionate by the medical staff, producing this seemingly counterintuitive effect. Regarding the variance of the frailty distribution, the cluster with the largest $\hat{\theta}_g$ is the first one, meaning that the hospital effect is stronger on patients in this cluster.

4 Conclusions

In this work, we introduce a novel parametric method for clustering data in the survival context, with an additional level of heterogeneity in the form of the frailty term. Our method can be inserted in the framework of CWM models, that we extend to time-to-event data. The proposed method allows to identify latent clusters of units that differ in terms of random variables and within which the parameters of the frailty survival model vary. This leads to the identification of heterogeneous survival dynamics. Through the different components of the model, we estimate the time-to-event with a great level of customization in the covariates and frailty section. The proposed method has been applied on a real case study to study the survival time of HF patients affected by COVID-19 in Lombardy region. Results reveal the presence of three patients profiles that differ in terms of age, gender and MCS and for whom the effects of specific respiratory diseases and of the facilities is different. This method contributes to the literature of the precision medicine and tailored treatments.

Acknowledgements. This work is part of the ENHANCE-HEART project: Efficacy evaluationN of the therapeutic-care patHways, of the heAlthcare providers effects aNd of the risk stratifiCation in patiEnts suffering from HEART failure. Andrea Cappozzo acknowledges financial support from the Italian Ministry of University and Research (MUR) under the Department of Excellence 2023–2027 grant agreement “Centre of Excellence in Economics and Data Science” (CEEDS). The present research is part of the activities of “Dipartimento di Eccellenza 2023–2027”.

References

1. Berta, P., Ingrassia, S., Punzo, A., Vittadini, G.: Multilevel cluster-weighted models for the evaluation of hospitals. *METRON* **74**(3), 275–292 (2016). <https://doi.org/10.1007/s40300-016-0098-3>
2. Berta, P., Vinciotti, V.: Multilevel logistic cluster-weighted model for outcome evaluation in health care. *Stat. Anal. Data Mining ASA Data Sci. J.* **12**(5), 434–443 (2019)
3. Bordes, L., Chauveau, D.: Stochastic EM algorithms for parametric and semiparametric mixture models for right- censored lifetime data. *Comput. Stat.* **31**, 1513–1538 (2016)
4. Kleinbaum, D.G., Klein, M.: *Survival Analysis a Self-Learning Text*. Springer, New York (1996)
5. Munda, M., Rotolo, F., Legrand, C.: parfm: Parametric frailty models in R. *J. Stat. Softw.* **51**, 1–20 (2012)
6. R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2021). <https://www.R-project.org/>
7. Vaupel, J.W., Manton, K.G., Stallard, E.: The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**(3), 439–454 (1979)



Random Effects in Unconditional Quantile Regressions

Luca Merlo^{1(✉)}, Nicola Salvati², Enrico Fabrizi³, Paolo Frumento⁴,
and Lea Petrella⁵

¹ Department of Human Sciences, European University of Rome, Rome, Italy
luca.merlo@unier.it

² Department of Economics and Management, University of Pisa, Pisa, Italy
nicola.salvati@unipi.it

³ Department of Business and Social Sciences, Università Cattolica del Sacro Cuore,
Piacenza, Italy
enrico.fabrizi@unicatt.it

⁴ Department of Political Sciences, University of Pisa, Pisa, Italy
paolo.frumento@unipi.it

⁵ MEMOTEF Department, Sapienza University of Rome, Rome, Italy
lea.petrella@uniroma1.it

Abstract. In this work we introduce a new unconditional quantile mixed model extending the Recentered Influence Function (RIF) regression approach for cross-sectional analysis to clustered data structures. To assess the effect of changes in the covariates on the quantiles of the unconditional distribution of the response in each group, we propose to estimate a random effects model for the conditional mean of the RIF on the explanatory variables. In the empirical analysis, we use a sample of US male workers from the 1983 Outgoing Rotation group supplement of the Current Population Survey to investigate the impact of unions on the distribution of hourly wages.

Keywords: Clustered data · influence function · linear mixed model

1 Introduction

Quantile Regression (QR) has attracted considerable interest in many empirical studies since its introduction in the seminal paper of [8]. It provides a way to model the conditional quantiles of a response as a function of explanatory variables in order to have a more complete picture of the entire conditional distribution compared to the classical mean regression. However, if one is interested in how the unconditional (marginal) distribution of the outcome responds to changes in the covariates, QR methods would yield misleading inferences (see [1, 2, 9]), as conditional effects will depend on the level of all other conditioning variables in the model. For example, the coefficient β_τ from a τ -th conditional quantile regression, $\beta_\tau = F_Y^{-1}(\tau | X = 1) - F_Y^{-1}(\tau | X = 0)$, is generally different

from the unconditional effect, $(\Pr[Y > q_\tau \mid X = 1] - \Pr[Y > q_\tau \mid X = 0]) / f_Y(q_\tau)$, on the τ -th quantile of the unconditional distribution of Y , q_τ . Motivated by this interest, [2] proposed the Unconditional Quantile Regression (UQR) for estimating the impact of changes in the distribution of the explanatory variables on quantiles of the unconditional distribution of a dependent variable. This method builds upon the concept of Recentered Influence Function (RIF) which originates from a widely used tool in robust statistics, namely the Influence Function (IF, see [5]). The RIF of a distributional statistic, say, ν is obtained by adding back the statistic to the IF and it can be thought of as the contribution of an individual observation on ν . In a regression framework with covariates, the UQR of [2] consists of running a mean regression of the RIF of the unconditional quantile of the outcome variable on the explanatory variables.

In many studies observations are often correlated with each other across time, space, or other dimensions, like groups. Frequently adopted sampling designs, including clustered, multilevel, spatial, and repeated measures, may induce dependency patterns between observations that the analysis needs to take into due account [4]. The RIF regression approach of [2], however, does not allow for the consideration of potential grouped data.

In this work, we thus extend the UQR to the case of clustered data by modeling the RIF using random effects models. By using the properties of influence functions, we show that the effect of a small location shift in the distribution of covariates on the unconditional quantiles of each group is given by the conditional average derivative of the RIF given the covariates and the random effects, producing the Unconditional Quantile Mixed Partial Effect (UQMPE). The UQMPE can be estimated by obtaining the RIF and then fitting random intercept models for the conditional mean of the RIF on the covariates.

2 Methods

Let Y denote a scalar random variable with absolutely continuous distribution function F_Y and let $\mathbf{X}$ denote the k -dimensional vector of explanatory variables. We consider a clustered data structure, in which some group or cluster effect U is present and denote with $q_{\tau u}$ the τ -th quantile of Y given $U = u$, i.e., $q_{\tau u} = \inf_q \{q : F_{Y|U}(q|u) \geq \tau\}$. By definition, the distribution function of Y conditional on U can be written as:

$$F_{Y|U}(y|u) = \int F_{Y|\mathbf{X},U}(y|\mathbf{X} = \mathbf{x}, U = u) dF_{\mathbf{X}}(\mathbf{x}). \quad (1)$$

To study the effect of an infinitesimal contamination in the underlying distribution $F_{Y|U}$ on $q_{\tau u}$, we can use the IF measure as in [2]. Let us consider Δ_y the probability measure that puts mass 1 at the value y and let $F_{Y,t\Delta_y} = (1-t)F_{Y|U} + t\Delta_y$ represent the mixing distribution with $t \in [0, 1]$. The IF of $q_{\tau u}$ at a point y is defined as

$$IF(y; q_{\tau u}, F_{Y|U}) = \lim_{t \rightarrow 0} \frac{q_{\tau u}(F_{Y,t\Delta_y}) - q_{\tau u}(F_{Y|U})}{t}. \quad (2)$$

Then, the RIF is defined as the first two terms of the von Mises linear approximation [11] of $q_{\tau u}(F_{Y|t\Delta_y})$ with $t = 1$, namely

$$RIF(y; q_{\tau u}, F_{Y|U}) = q_{\tau u} + \int IF(s; q_{\tau u}, F_{Y|U}) d\Delta_y(s) = q_{\tau u} + IF(y; q_{\tau u}, F_{Y|U}). \quad (3)$$

In presence of covariates $\mathbf{X}$ and group effects U , by using the law of iterated expectations we can express $q_{\tau u}$ in terms of the conditional expectation of $RIF(y; q_{\tau u}, F_{Y|U})$ given $\mathbf{X}$ and U :

$$q_{\tau u} = \int RIF(y; q_{\tau u}, F_{Y|U}) dF_{Y|U}(y|u) = \int \mathbb{E}[RIF(y; q_{\tau u}, F_{Y|U}) | \mathbf{X} = \mathbf{x}, U = u] dF_{\mathbf{X}}(\mathbf{x}), \quad (4)$$

where the first equality follows from the fact that the IF integrates to zero [6], and the second equality come from substituting in (1). Expression (4) shows that the impact of covariates on $q_{\tau u}$ can be obtained integrating over

$\mathbb{E}[RIF(y; q_{\tau u}, F_{Y|U}) | \mathbf{X} = \mathbf{x}, U = u]$, which is easily done using mixed regression models. Here, our objective is to identify how a small location shift in the distribution of $\mathbf{X}$ affect the quantile of the unconditional distribution of $F_{Y|U}$ given some group effects U . From (4) it follows that the vector of these unconditional effects is given by:

$$\boldsymbol{\alpha}(\tau|u) = \int \frac{d\mathbb{E}[RIF(\mathbf{Y}; q_{\tau u}, F_{Y|U}) | \mathbf{X} = \mathbf{x}, U = u]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}), \quad (5)$$

where $d\mathbb{E}[RIF(\mathbf{Y}; q_{\tau u}, F_{Y|U}) | \mathbf{X} = \mathbf{x}, U = u]/d\mathbf{x}$ is understood to indicate the $k \times 1$ vector of partial derivatives with respect to $[\mathbf{x}_j]_{j=1}^k$. We call the quantity $\boldsymbol{\alpha}(\tau|u)$ in (5) as the Unconditional Quantile Mixed Partial Effect (UQMPE). By analogy with standard conditional regression coefficients, $\boldsymbol{\alpha}(\tau|u)$ corresponds to the effect of a small increase in the location of the distribution of the explanatory variables on $q_{\tau u}$, holding everything else constant.

3 Estimation

We denote by $j = 1, \dots, J$ the index of the cluster and by $i = 1, \dots, n_j$ the within-cluster index, such that the total sample size is $n = \sum_{j=1}^J n_j$. Let Y_{ij} and $\mathbf{X}_{ij}$ be, respectively, the response variable and vector of covariates for the i -th observation in cluster j . Also, let $\mathbf{X}_j$ be the $n_j \times k$ design matrix for cluster j . For a given τ , the first step of the analysis requires to estimate $q_{\tau j}$ and f_{Y_j} . To make efficient use of the available information and obtain stable estimates, we assume that $q_{\tau j}$ and f_{Y_j} are common to all clusters, i.e., $q_{\tau j} = q_{\tau}$ and $f_{Y_j} = f_Y$, $\forall j = 1, \dots, J$. To estimate q_{τ} we suggest to proceed using the parametric quantile regression coefficients modeling (QRCM) framework described in [10], and then estimate f_Y using the corresponding quantile density function. Finally, to estimate the UQMPE, we assume that the group-specific RIF can be modeled using a simple random-intercept model that can be written as:

$$RIF(Y_{ij}; q_{\tau}, F_Y) = \mathbf{x}'_{ij} \boldsymbol{\beta}_{\tau} + u_{\tau j} + e_{\tau ij}, \quad j = 1, \dots, J, \quad i = 1, \dots, n_j, \quad (6)$$

where $e_{\tau ij}$ is the τ -specific individual error term.

4 Application

As an example of the proposed method, we investigate the impact of unions on wages, which is well known to be different at different points of the wage distribution. We use a sample of 1707 observations on US males from the 1983 Outgoing Rotation Group supplement of the Current Population Survey. The dependent variable is the real log hourly wage for all wage and salary workers, and the explanatory variables include dummy variables for unionized, married and black workers, and continuous ones for years of education and years of labor market experience. We fit the proposed model at quantile levels $\tau = \{0.1, 0.5, 0.9\}$, with groups specified for job class indicating type of job, and compare the results with the (conditional) linear quantile mixed model (LQMM) of [4]. Table 1 reports the results with point estimates displayed in boldface when significant at the 5% level. As can be shown, unions have a positive and statistically significant effect at all three quantile levels, as evaluated by the LQMM. On the contrary, the UQMPE initially rises from 0.410 at the 10-th quantile to 0.525 at the median before falling (-0.440) at the 90-th quantile. These findings support the point made earlier that unions have distinct effects at different parts in the wage distribution.

Table 1. Estimated regression coefficients at $\tau = \{0.1, 0.5, 0.9\}$ for linear quantile mixed models and unconditional quantile mixed regressions.

Variable	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
	LQMM	UQMR	LQMM	UQMR	LQMM	UQMR
Intercept	-3.023 (0.275)	-3.680 (0.263)	-3.315 (0.143)	-3.930 (0.263)	-2.673 (0.233)	-2.746 (0.345)
Union Status	0.473 (0.119)	0.303 (0.058)	0.289 (0.074)	0.659 (0.099)	0.230 (0.103)	-0.491 (0.121)
Non-white	-0.125 (0.068)	0.064 (0.106)	-0.135 (0.061)	-0.194 (0.102)	-0.128 (0.149)	-0.087 (0.131)
Married	0.279 (0.074)	0.248 (0.075)	0.264 (0.046)	0.395 (0.075)	0.170 (0.070)	0.078 (0.108)
Education	0.119 (0.015)	0.126 (0.016)	0.194 (0.011)	0.230 (0.016)	0.221 (0.011)	0.261 (0.023)
Experience	0.013 (0.003)	0.021 (0.004)	0.027 (0.002)	0.024 (0.003)	0.029 (0.004)	0.025 (0.003)

5 Conclusion

In this paper, we propose a new linear mixed regression to model unconditional quantiles exploiting the concept of RIF for clustered data structures. To evaluate the influence of covariate changes on the quantiles of the unconditional distribution of the response in each group and to explain between-group variance, we develop a two-step procedure. In the first step, we estimate the unconditional quantile of the response variable common to all clusters using the QRCM framework in [10]. In the second one, we fit a random effects model for the conditional mean of the RIF on the explanatory variables. We present an application that investigates the effect of unions on the wage distribution. While traditional linear mixed quantile regression indicates that unions have a positive effect on wages even at the top quantiles of the unconditional wage distribution, we find a strong negative effect of unions at the highest quantile of the wage distribution.

References

1. Borah, B.J., Basu, A.: Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. *Health Econ.* **22**(9), 1052–1070 (2013)
2. Firpo, S., Fortin, N.M., Lemieux, T.: Unconditional quantile regressions. *Econometrica* **77**(3), 953–973 (2009)
3. Frumento, P., Bottai, M., Fernández-Val, I.: Parametric modeling of quantile regression coefficient functions with longitudinal data. *J. Am. Stat. Assoc.* **116**(534), 783–797 (2021)
4. Geraci, M., Bottai, M.: Linear quantile mixed models. *Stat. Comput.* **24**, 461–479 (2014)
5. Hampel, F.R.: The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **69**(346), 383–393 (1974)
6. Rousseeuw, P. J., Hampel, F. R., Ronchetti, E. M., Stahel, W. A.: *Robust statistics: the approach based on influence functions*. John Wiley & Sons (2011)
7. Koenker, R.: Quantile regression for longitudinal data. *J. Multivar. Anal.* **91**(1), 74–89 (2004)
8. Koenker, R., Bassett Jr, G.: Regression quantiles. *Econometrica* 33–50 (1978)
9. Merlo, L., Petrella, L., Salvati, N., Tzavidis, N.: Unified unconditional regression for multivariate quantiles, M-quantiles, and expectiles. *J. Am. Stat. Assoc.* 1–12 (2023)
10. Frumento, P., Bottai, M.: Parametric modeling of quantile regression coefficient functions. *Biometrics* **72**(1), 74–84 (2016)
11. Mises, R.V.: On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **18**(3), 309–348 (1947)



Teachers' Evaluations and Ethnic Penalty on Academic Achievement in Italy

Eleonora Miaci¹, Giulia Berta¹(✉), Martina Sciarabba², and Nazareno Panichella¹

¹ Università degli Studi di Milano, 7, 20122 Milano, Italy
{eleonora.miaci,nazareno.panichella}@unimi.it,
giuliaberta97@gmail.com

² Utrecht University 8, 3584 Utrecht, The Netherlands

Abstract. This study, utilizing 2018–2019 INVALSI data, aims to bridge the knowledge gap regarding the school experiences of students from migratory backgrounds in Italy. Its primary aim is to investigate the penalization in teacher evaluations faced by middle school students with migration backgrounds, compared to their native counterparts. Employing Ordinary Least Squares (OLS) regressions, the analysis examines disparities existing between anonymously graded INVALSI test scores and first-semester grades that may be influenced by teacher biases. Findings bring to light a systematic underassessment of migratory students, especially pronounced for those of non-European descent. Moreover, variations in these disparities are discerned across different migratory generations and the parental combinations of national origins. Even after controlling for gender, social origin, and academic performance, the ethnic disparity persists, highlighting the impact of underlying biases and systemic factors. The insights resulting from the analysis encourage the implementation of policy interventions aimed to bolster educational equity and inclusivity.

Keywords: Educational Disparities · Migratory Generations · INVALSI Test Scores

1 Theoretical Background

Empirical research emphasizes the central role of schools in social stratification processes. Hence, the academic success of second generations can be a valid indicator of the level of integration of the immigrant population (Ballarino & Panichella 2021). Previous studies have highlighted ethnic differences in educational performances between students with and without a migration background (Heath et al., 2008; Alba & Foner 2015). Empirical research focused on different mechanisms of penalization, such as peer effect, language barriers, home environment, cultural capital match/mismatch, oppositional culture and teacher bias (Kozlowski 2015; Ballarino & Panichella 2021).

The impact of teachers' assessments on students' performances has been neglected, despite its critical importance in explaining Ethnic Educational Disadvantage and the perpetuation of educational inequalities (Argentin and Pavolini 2020). Teachers' grading plays a pivotal role in shaping students' life choices, such as the decision on further educational tracks to pursue or the level of investment their families are willing to make in their education (Ballarino & Panichella 2021). Moreover, it significantly influences their motivation and engagement, often leading to a "self-fulfilling prophecy" where expectations directly impact outcomes (Chandrasegaran, 2018).

School evaluations are influenced by more than just academic achievements (Dee 2005). Research has shown that teachers' biases, shaped by factors like students' migration backgrounds and gender, can significantly affect grading (Malouff & Thorsteinsson 2016). Critical educational theories emphasize that, despite teachers' intentions towards objectivity and inclusivity, there is a tendency to favour students whose behaviours and characteristics reflect their own (Zee & Koomen 2017).

Most recent research has specifically focused on the role of gender and ethnicity, asking whether teachers' evaluations systematically change between male and female and among students with different ethnic backgrounds.

International research indicates that teachers globally tend to assess students from marginalized social groups more negatively than their advantaged counterparts, beyond objective performance metrics. For instance, analyses in Italy (Triventi, 2020), New Zealand (Meissel et al., 2017), Sweden (Hinnerich et al., 2015), and the UK (Burgess & Greaves, 2013) have shown systematic undervaluation of students from marginalized ethnic groups compared to advantaged groups, regardless of actual performance or in comparison with blind evaluators' assessments.

In Germany, mixed results emerge from the literature: while some research reveals that teachers' biases lead to lower grades for students with perceived migrant backgrounds (Sprietsma, 2013), other studies find no evidence of grade discrimination against secondary school immigrants (Kiss, 2013). In the Netherlands, Van Ewijk (2011) identified no direct grading bias by ethnic majority teachers against minority students but uncovers indirect biases through lower teacher expectations and unfavourable attitudes that may lead to minority students underperforming.

2 The Aim of This Study

This study aims to contribute to the ongoing debate on ethnic penalties in education by focusing on the role of teachers' assessments in shaping these disparities within the Italian context. Italy presents a particularly relevant case study due to its status as a country of recent migration, where there exists a notable gap in knowledge concerning the experiences of students with a migration background in the school context compared to other European countries.

Our primary objective is to investigate whether students with a migration background are systematically penalized by teachers' evaluations compared to their native peers, and how this varies across different migratory generations.

The second aim of this study is to address the literature gap in the interplay between gender and ethnicity, examining whether ethnic penalties in teachers' evaluations are lower among female students.

Thus, we aim to contribute by shedding light on the double penalties of ethnicity and gender and on their implications for the cumulative disadvantages in the life courses of migrants.

3 Data and Analysis

To test our hypothesis, we draw on 2018–2019 data from the INVALSI surveys (National Institute for the Evaluation of the Educational System), which assess the Italian student population's educational performance. Our analysis specifically focuses on the results of middle school students in mathematics tests. Within this dataset, a total of 541,297 observations have been compiled. Among these, 76.26% pertain to students of Italian origin, whereas the remainder is composed of students with varying migratory backgrounds.

We estimated a set of OLS regression models to explore the discrepancy between scores from INVALSI tests, which are graded anonymously, and first-semester grades that might be influenced by teacher bias. The empirical strategy consists of three steps. The first step examines the entire sample and assesses the ethnic gap in academic achievement. It investigates whether students with varying migratory backgrounds (classified as 0 for native, 1.5 generation, 2nd generation, and 2nd generation with one native parent and one migrant parent) are disadvantaged compared to native students.

The second step analyzes whether these differences persist significantly even when considering teachers' evaluations as a control. It investigates whether there is a migration penalty on academic achievement measured by INVALSI tests even when considering teacher assessments.

The third step explores whether the ethnic penalty on educational achievement varies among students with different teacher evaluations. Specifically, it investigates how the ethnic penalty on INVALSI scores changes among students who receive different ratings from their teachers.

4 Primary Results

Table 1 presents the preliminary results of our OLS regressions. Models 1 and 2 compare first-semester academic outcomes across students from diverse migratory generations. Model 1 does not include control variables, while Model 2 controls for gender and social origin. The data enable a detailed assessment by considering parents' education and profession levels. The findings indicate that students from various migratory backgrounds are at a disadvantage compared to native students. Notably, students from mixed-parentage households exhibit different patterns of disadvantage. For families with an Italian father and a migrant mother, the disadvantage persists, yet it diminishes when the parent genders are reversed. To interpret our findings, we propose that mothers are generally more involved in their children's education compared to fathers, and an Italian mother could provide her children with greater linguistic and cultural capital directly relevant to the Italian educational context. This may encompass not just the language,

but also an understanding of cultural values, expectations, and behaviors that are positively regarded in a school environment. Additionally, mothers in Italy tend to be more engaged in the educational process, including interaction with teachers and participation in school life. The presence of an Italian mother could somewhat mitigate unconscious biases held by teachers, who might perceive students as more closely aligning with their cultural expectations. In Model 3, we add INVALSI math scores to the previous controls. The analysis indicates that migrant students still face a penalty in school evaluations, even when accounting for INVALSI results, suggesting that differences cannot be solely attributed to objectively lower performance. Model 4 then investigates the interaction between migratory background and INVALSI scores, examining whether the disadvantage for migrant students is more pronounced among those with higher or lower academic performance. The data suggest that the ethnic disadvantage is pervasive and does not discriminate based on academic merit, implying that external factors such as potential biases could significantly influence teacher evaluations. Furthermore, this effect predominantly affects students originating from non-European countries.

Figure 1 synthesizes our findings by illustrating the disparities in academic performance among students from different migratory generations and how these disparities remain consistent across the entire distribution of INVALSI test scores.

Table 1. OLS regressions on first-semester grades. Data: INVALSI survey 2018–2019

First-semester grades	MODEL 1	MODEL 2	MODEL 3	MODEL 4
Migratory gen (ref. Italian)				
1.5 generation	−0.65*** (−0.75–0.55)	−0.35*** (−0.46–0.25)	−0.28*** (−0.36–0.19)	−0.35*** (−0.58–0.12)
2.generation	−0.58*** (−0.63–0.52)	−0.33*** (−0.39–0.27)	−0.33*** (−0.38–0.28)	−0.33*** (−0.48–0.18)
Ita father, migrant mother	−0.14*** (−0.21–0.06)	−0.11*** (−0.18–0.03)	−0.13*** (−0.19–0.06)	−0.20* (−0.42–0.03)
Ita mother, migrant father	0.01 (−0.08–0.11)	0.02 (−0.08–0.11)	−0.07* (−0.15–0.01)	−0.08 (−0.37–0.22)
Europeans	−0.23*** (−0.27–0.19)	−0.13*** (−0.17–0.09)	−0.13*** (−0.16–0.10)	−0.08 (−0.18–0.02)
Other	−0.38*** (−0.40–0.36)	−0.21*** (−0.24–0.19)	−0.17*** (−0.20–0.15)	−0.21*** (−0.27–0.15)

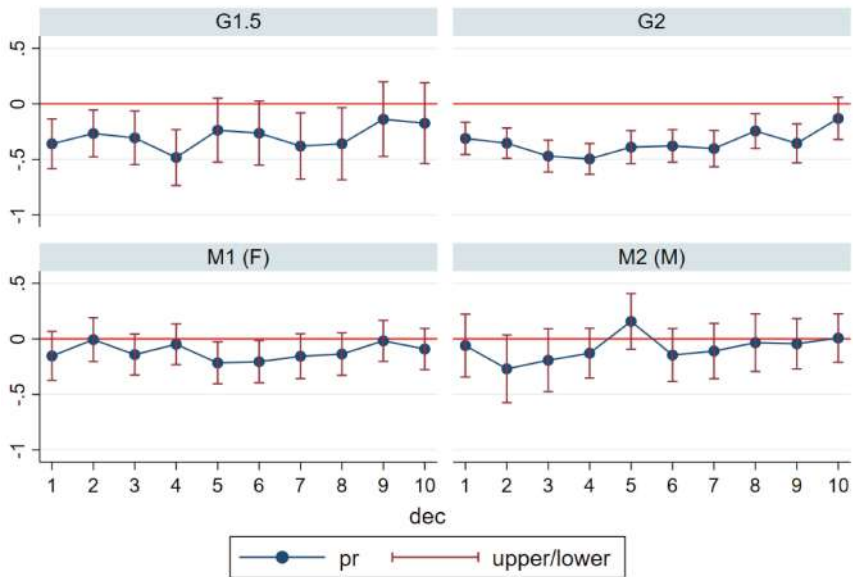


Fig. 1. First-semester grades by migratory generation across INVALSI deciles. Data: INVALSI survey 2018–2019.

5 Conclusions

Our analysis, based on the 2018–2019 INVALSI data, confirms the existence of educational achievement disparities among middle school students in Italy. Through a sequence of OLS regressions, we have identified significant impacts of migratory generations on academic outcomes, even when controlling for gender and social origin. Notably, our findings highlight the enduring disadvantage experienced by students from migratory backgrounds, especially those from non-European countries. This disadvantage persists even when accounting for standardized test scores, indicating that it cannot be solely attributed to objective performance deficits. The distinct patterns observed in households with mixed parentage emphasize the continuing significant role played by mothers, particularly within the Italian context, in their children's education. The presence of an Italian mother appears to confer linguistic and cultural advantages relevant to the educational setting, including language proficiency and an understanding of cultural values. Moreover, mothers in Italy are typically more engaged in the educational process, from interacting with teachers to participating in school life, potentially mitigating unconscious biases teachers may hold, leading to a perception of students as more aligned with their cultural expectations. Moreover, the ethnic disadvantage's pervasive nature, unaffected by academic merit, points to the potential influence of biases, deep-rooted within the educational system, in determining these outcomes. These insights underscore the need to reassess educational policies and practices towards greater inclusiveness and equity.

References

- Alba, R., Foner, N.: *Strangers No More. Immigration and Challenges of Integration in North America and Western Europe*. Princeton University Press, Princeton and Oxford (2015)
- Argentin, G., Pavolini, E.: How schools directly contribute to the reproduction of social inequalities. Evidence of tertiary effects, taken from Italian research. *Soc. Policies* **7**(1), 149–176 (2020)
- Ballarino, G., Panichella, N.: *Sociologia dell'istruzione. Il mulino* (2021)
- Burgess, S., Greaves, E.: Test scores, subjective assessment, and stereotyping of ethnic minorities. *J. Labor Econ.* **31**, 535–576 (2013)
- Chandrasegaran, J.: The role of self-fulfilling prophecies in education: teacher-student perceptions. *J. Educ. Psychol.* **12**(1), 8–18 (2018)
- Dee, T.S.: A teacher like me: does race, ethnicity, or gender matter? *Am. Econ. Rev.* **95**(2), 158–165 (2005)
- Heath, A.F., Rethon, C., Kilpi, E.: The second generation in Western Europe: education, unemployment, and occupational attainment. *Ann. Rev. Sociol.* **34**, 211–235 (2008)
- Hinnerich, B.T., et al.: Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Educ. Econ.* **23**, 660–676 (2015)
- INVALSI: *Rapporto nazionale sulla rilevazione degli apprendimenti 2011–12*. Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione, Roma (2012)
- Kozlowski, K.P.: Culture or teacher bias? racial and ethnic variation in student-teacher effort assessment match/mismatch. *Race Soc. Probl.* **7**, 43–59 (2015)
- Kiss, D.: Are immigrants and girls graded worse? Results of a matching approach. *Educ. Econ.* **21**(5), 447–463 (2013)
- Malouff, J.M., Thorsteinsson, E.B.: Bias in grading: a meta-analysis of experimental research findings. *Aust. J. Educ.* **60**(3), 245–256 (2016)
- Meissel, K., et al.: Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ.* **65**, 48–60 (2017)
- Sprietsma, M.: Discrimination in grading: experimental evidence from primary school teachers. *Empir. Econ.* **45**, 523–538 (2013)
- Triventi, M.: Are children of immigrants graded less generously by their teachers than natives, and why? Evidence from student population data in Italy. *Int. Migr. Rev.* **54**(3), 765–795 (2020)
- Van Ewijk, R.: Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Econ. Educ. Rev.* **30**(5), 1045–1058 (2011)
- Zee, M., Koomen, H.M.: Similarities and dissimilarities between teachers' and students' relationship views in upper elementary school: the role of personal teacher and student attributes. *J. Sch. Psychol.* **64**, 43–60 (2017)



Improving Interpretability Using Surrogate Models: An Ecological Case Study

Gianni Gilioli , Manlio Migliorati  , and Anna Simonetto 

Department Civil, Environmental, Architectural Engineering and Mathematics (DICATAM),
University of Brescia, via Branze 43, 25123 Brescia, Italy
`manlio.migliorati@unibs.it`

Abstract. Understanding of machine learning models is crucial for Artificial Intelligence acceptance, particularly for domains focused on highly interconnected and biological processes. In this manuscript, our experience in improving readability of Random Forest regression results via surrogate models, i.e. models trying to replicate non transparent fitting results in an interpretable framework, is reported. A Random Forest regression model describing the abundance of *Anopheles* mosquitoes in the northern Italy has been implemented. Then, surrogate models have been built and analysed to improve readability of the original model. Adequacy of surrogate models is stated on the basis of both classical coefficient of determination R^2 and Variable Importance ranking comparison.

Keywords: Machine Learning · Interpretability · Surrogate Models · Variable Importance

1 Introduction

Machine learning methods should not only provide good fitting results, but also offer interpretability mechanisms helping users in better understanding a model. The eXplainable Artificial Intelligence (XAI) [1] research field is becoming more and more important, analysing how to integrate good fitting algorithms and satisfactory explanation mechanisms.

Some machine learning algorithms (for instance Linear Regression or Decision Trees [2]) are transparent, providing not only Variable Importance ranking, but coupling them with other indicators supporting the interpretability of a model, as coefficient sign or visual representation.

Some other algorithms, as the ensemble learning [3] Random Forest algorithm [4], can be defined opaque, offering interpretation tools limited to Variable Importance because of their nature based on considering a huge number of models.

Other machine learning approaches, such as SVM (Support Vector Machine) [5], Neural Networks [6], Deep Learning [7] or Boosting methods [8] are considered as black box methods: the model results must be taken as they are, without the possibility of understanding the mechanism generating them.

In this work, the good fitting/low interpretability methods have been coupled to easily interpretable approaches, to build the so-named ‘surrogate models’ [1] helping in understand fitting results.

The key concept is simple: a surrogate model tries to replicate the predictions of an opaque or black box model (hereafter original model) applying an interpretable mechanism to the same regressors values. If the original and the surrogate models are similar enough, it will be possible to use the additional information of the surrogate models to better understand the original results.

In this manuscript, our experience in building surrogate models related to a Random Forest model fitted on a dataset collecting information about the abundance of *Anopheles* mosquitoes in northern Italy [9] is reported. Surrogate models have been built using both Linear Regression and Decision Trees.

All implementations have been made using the *R* language [10] in the *RStudio* environment [11].

2 Surrogate Models

Surrogate models are built using easily interpretable machine learning algorithms (e.g. Linear Regression, Decision Trees) working on predictions (the original model) obtained by machine learning approaches typically offering a high fitting quality but few or no tools to understand results.

Surrogate models are not related to real values of the dependent variable, but only to its predictions included in the original model.

The objective is to minimize discrepancy between the surrogate model’s predictions and the predictions of the original model.

More formally: let’s suppose we observe a quantitative (we are in a regression context) dependent variable Y , and p regressors ($X_1, X_2, \dots, X_p$). The regression problem can be expressed [12] as

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon \quad (1)$$

where f is an unknown function and ε is the error term.

Let’s suppose that a prediction vector $\hat{Y}$ has been estimates for Y applying an opaque or black box model:

$$\hat{Y} = \hat{f}(X_1, X_2, \dots, X_p) \quad (2)$$

where $\hat{f}$ is an estimate for the function f .

A surrogate model tries to emulate predictions for vector $\hat{Y}$ using a new function $\hat{g}$, applied to the same regressors vector as in the original problem, to produce $\hat{\hat{Y}}$, a vector of predictions for predictions of the original model:

$$\hat{\hat{Y}} = \hat{g}(X_1, X_2, \dots, X_p) \quad (2)$$

To understand if the two functions $\hat{f}$ and $\hat{g}$ are close enough, in a regression context it is possible to analyse the coefficient of determination R^2 of the surrogate model, to compare its fitting quality with respect to the original model.

Moreover, if the original model provides Variable Importance measures (for instance when it is generated by Random Forest), it can be useful to compare variables ranking between the surrogate model and the original one, verifying consistency among Variable Importance rankings.

Surrogate models can be a valuable tool in machine learning because they enable clearer and faster explorations and analysis of complex models, especially when dealing with computationally expensive datasets with many features.

To this purpose, well known techniques offering a good interpretability as Linear Regression (providing several information about regressors by means of magnitude and sign coefficients) or Decision Trees (providing a simple and appealing graphical representation) can be used.

3 Case Study: The Abundance of the *Anopheles* Mosquitoes in Northern Italy

3.1 The Problem and Its Dataset

The case study concerns the environmental factors (land use and meteorological variables) mostly influencing the presence of *Anopheles* mosquitoes in Lombardy and Emilia Romagna regions (Northern Italy) [13]. These two regions were characterized in the past by high population abundance of *Anopheles* mosquitoes, and they still present environmental conditions locally suitable for these malaria vectors.

The dataset includes 858 mosquito samplings conducted in the study area from 2017 to 2020. Each sampling is described through 25 variables:

- the $\log_{10}$ abundance of *Anopheles* captured in a sampling (the dependent variable)
- four meteorological variables related both to temperature and rainfall, and twenty land use variables (the regressors).

All observations are complete: in the dataset there are no missing values.

3.2 Random Forest Model

Random Forest algorithm has been applied to the original dataset, to predict log-abundance of *Anopheles* mosquitoes. The produced model, fitted on the base of 1000 trees, has a R^2 equal to 0.43, acceptable in the ecological context of this case study.

Figure 1 reports the ranking of Variable Importance of the Random Forest model, sorted in descending order, both considering the percentage Increase in Mean Squared Error ($\%IncMSE$) and the Increase in Node Purity ($IncNodePurity$).

To make easier comparisons, values have been normalized: the most important regressor value has been set to 100, and values of other regressors have been accordingly resized. The variable importance can be useful to interpret the model, but it is not enough for case study purposes.

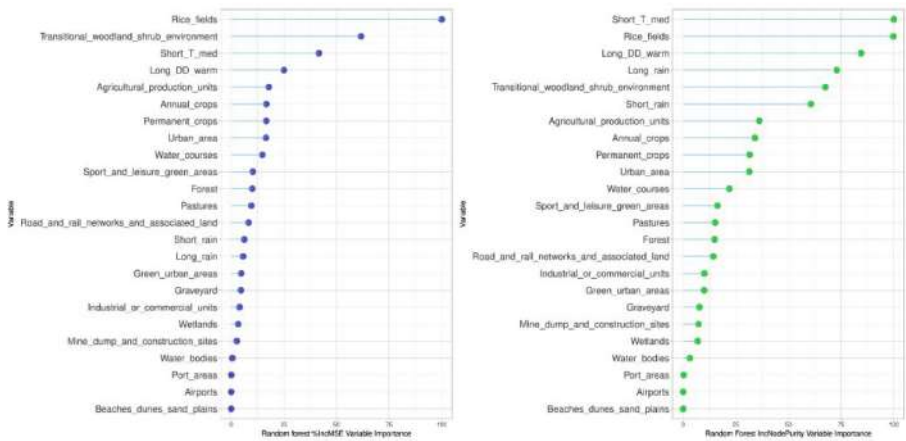


Fig. 1. Original Random Forest model: scaled (to 100) Variable Importance, calculated using both *%IncMSE* (on the left) and *IncNodePurity* (on the right). Regressor names are on y-axis, scaled Variable Importance on x-axis.

3.3 Linear Regression Surrogate Model

The first potential surrogate model was built using Linear Regression. Unfortunately, the generated model is not similar enough to the original one: it exhibits a low value for R^2 index (equal to 0.44), typically addressing nonlinearity of data in the dataset, and it cannot be used as surrogate for the original Random Forest model.

3.4 Decision Tree Surrogate Model

The second potential surrogate model is a Decision Tree, built via the *R* rpart package [13].

The Decision Tree surrogate model exhibits a R^2 index equal to 0.79 and its ranking of Variable Importance is close to the ranking of the original mode. We conclude that the fitted Decision Tree can be used for improving readability of the original Random Forest model.

Figure 2 shows the structure of the Decision Tree surrogate model. The graphical representation shows that the first split in the tree surrogate model is based on the regressor ‘Rice field’ (expressing the percentage of rice fields in the area we are considering). The result is consistent with the Random Forest results, where ‘Rice field’ was the most important variable. Moreover, via the surrogate model we have other information: for instance, the split value (1.8) of this regressor is made available, too.

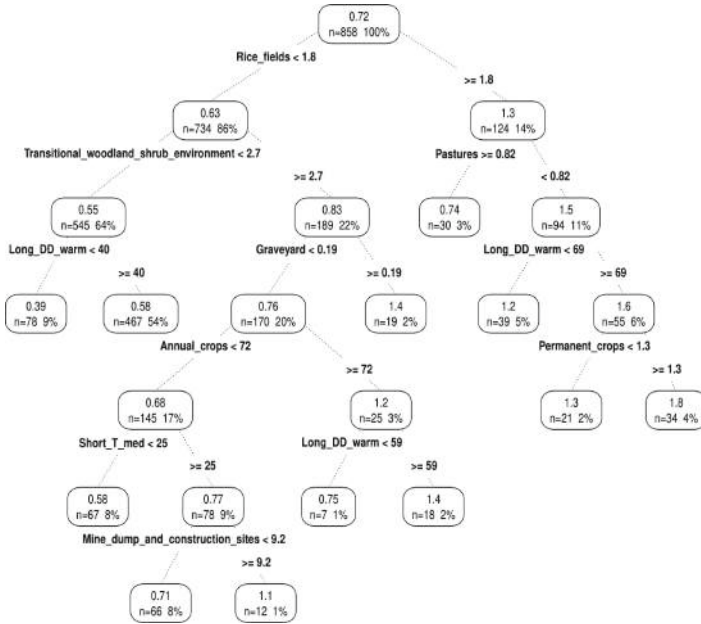


Fig. 2. Decision Tree surrogate model. Nodes shows the value of the dependent variable, the number of observations belonging to the node and its percentage with respect to the total number of observations. Arcs are marked with the regressor used in the split and its split value.

4 Conclusions

In this paper surrogate models facilitating interpretation of opaque models generated by Random Forest algorithm has been reported, for a case study with the goal of identifying the environmental factors (land use and local meteorological variables) mostly influencing the presence of *Anopheles* in the Lombardy and Emilia Romagna regions (Northern Italy).

The basic idea is simple and effective: it can be possible to use interpretable machine learning algorithms to better understand opaque/black box models, trying to estimate the predictions of the original models instead of the real dependent variable values.

In our case study Linear Regression surrogate model has a R^2 of 0.44, too low to be accepted. Instead, Decision Tree model not only has a R^2 equal to 0.79, but also its variable importance ranking is close to original variable importance values, therefore it can be used to have supplementary information about results.

The most important prerequisite to surrogate model application is to have an adequate fitting quality for the original model, avoiding building a useless high fitting quality surrogate model for a bad quality original model.

Next steps of this work will be focused on the verification of other algorithms that can be used for building interpretable meta-models, as for instance the conditional inference trees described in [14], and on the application of surrogate models to classification problems.

References

- Molnar, C.: Interpretable Machine Learning, 2nd edition. (2022). <https://christophm.github.io/interpretable-ml-book>
- Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Chapman and Hall/CRC, Boca Raton, Fla (1984)
- Dietterich, T.G.: Ensemble methods in machine learning. Multiple Classifier Syst. **1857**, 1–152 (2000). <https://doi.org/10.1007/3-540-45014-91>
- Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
- Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
- Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
- Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press, Boston (2016)
- Schapire, R.E., Freund, Y.: Boosting: Foundations and Algorithms. The MIT Press, Boston (2012)
- Gilioli, G., et al.: Characterization of environmental drivers influencing the abundance of *Anopheles maculipennis* complex in Northern Italy. Parasites & Vectors **17**, 109 (2024). <https://doi.org/10.1186/s13071-024-06208-6>
- R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. R Foundation for Statistical Computing (2023). <https://www.R-project.org/>
- RStudio Team: RStudio: Integrated Development Environment for R. (2023). <http://www.rstudio.com/>
- James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer (2013). <https://faculty.marshall.usc.edu/gareth-james/ISL/>
- Therneau, T., Atkinson, B.: rpart: Recursive Partitioning and Regression Trees (2022). R package version 4.1.19. <https://CRAN.R-project.org/package=rpart>
- Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. **15**(3), 651–674 (2006). <https://doi.org/10.1198/106186006X133933>



Statistical Analysis of Peer Grading: A Latent Variable Approach

Giuseppe Mignemi¹(✉), Yunxiao Chen², and Irini Moustaki²

¹ Bocconi Institute for Data Science and Analytics, Milan, Italy
giuseppe.mignemi@unibocconi.it

² London School of Economics, London, England
{Y.Chen186, I.Moustaki}@lse.ac.uk

Abstract. Peer grading is an educational system in which students assess each other's work. It is commonly applied under Massive Open Online Course (MOOC) and offline classroom settings. Peer grading data have a complex network structure, where each student is a vertex of the network, and each peer grade serves as an edge connecting one student as a grader to another student as an examinee. We introduce a latent variable model framework for analyzing peer grading data and develop a fully Bayesian procedure for its statistical inference. The proposed approach produces more accurate aggregated grades by modelling the heterogeneous grading behaviour with latent variables and provides a way to assess each student's performance as a grader. It may be used to identify a pool of reliable graders or generate feedback to help students improve their grading. Thanks to the Bayesian approach, uncertainty quantification is straightforward when inferring the student-specific latent variables as well as the structural parameters of the model. The proposed method is applied to a real-world dataset.

Keywords: peer grading · rating models · cross-classified models · latent variable approach · Bayesian modelling

1 Introduction

Peer grading, also known as peer assessment, is a system of formative assessment in education whereby students assess and give feedback on one another's assignments. It substantially reduces teachers' burden for grading and improves students' understanding of the subject and critical thinking [9]. Consequently, it is widely used in many educational settings, including massive open online courses (MOOCs; [4]), large university courses [2], and small classroom settings [7].

2 Proposed Model

Consider N students who receive T assignments. Each student i 's work on assignment t is randomly assigned to a small subset of other students to grade their work. We denote this subset as S_{it} , which is a subset of $\{1, \dots, i-1, i+1, \dots, N\}$. Each grader $g \in S_{it}$ gives

a grade Y_{igt} to this work, following certain scoring rubrics. For simplicity, we consider the case when Y_{igt} is continuous while pointing out that extending the proposed model to ordinal data may be achieved using an underlying variable formulation [1]. It is common, but not required, for the number of grades $|S_{it}|$ to be the same for all students and assignments. An aggregated score is then computed as a measure of student i 's performance on the t th assignment student i 's performance on the t th assignment, often by taking the mean or the median of the peer grades $Y_{igt}, g \in S_{it}$. We note that a simple aggregation rule, such as the mean and the median of the peer grades, fails to account for the grader effect and, thus, may not be accurate enough.

2.1 Proposed Model

Modelling Peer Grade Y_{igt} . We assume the following decomposition for the peer grade Y_{igt} :

$$Y_{igt} = \theta_{it} + \tau_{igt} - \delta_t, \quad i = 1, \dots, N, t = 1, \dots, T, g \in S_{it}. \quad (1)$$

Here, δ_t captures the difficulty level of assignment t . A larger value of δ_t corresponds to a more difficult assignment. In addition, θ_{it} represents student i 's true score for assignment t , and τ_{igt} is an error brought by the grader. We assume θ_{it} , τ_{igt} and δ_t to be independent.

Modelling True Score θ_{it} . For each student i , we assume that their true scores for different assignments θ_{it} , $t = 1, \dots, T$, are independent and identically distributed, following a normal distribution

$$\theta_{it} \sim N(\alpha_i, \eta_i^2), \quad (2)$$

where the mean and variance are student-specific latent variables. The latent variable α_i captures the student's average performance over the assignments, and the latent variable η_i^2 measures their performance stability. This model assumes the true scores fluctuate randomly around the average score α_i without a trend. This assumption can be relaxed if we are interested in assessing students' growth over time.

Modelling Grader Effect τ_{igt} . Each student g grades multiple assignments from multiple students. We let $H_g = \{(i, t) : g \in S_{it}, t = 1, \dots, T\}$ be all the work student g grades. For each student g , we assume that τ_{igt} , for all $(i, t) \in H_g$, are independent and identically distributed (i.i.d.), following a normal distribution $N(\beta_g, \phi_g^2)$, where the mean and variance are student-specific latent variables. The latent variable β_g may be interpreted as the bias of student g as a grader. For two students g and g' satisfying $\beta_g > \beta_{g'}$, student g will give a higher grade on average than student g' when grading the same work. We say grader g is unbiased when $\beta_g = 0$. Moreover, the latent variable ϕ_g^2 is a measure of grader reliability. A smaller value of ϕ_g^2 implies that the grader tends to follow a consistent standard, while a larger value suggests that they may give erratic grades that lack a consistent standard. In other words, when grading multiple pieces of work with the same true score and assignment difficulty (so that ideally they should receive the same grade), a grader with a small ϕ_g^2 tends to give similar grades, and thus, the grades are more reliable. In contrast, a grader with a large ϕ_g^2 tends to give noisy grades that lack consistency. We remark that the grader effects τ_{igt} , $t = 1, \dots, T$, are assumed to be i.i.d. in the current setting, which means the grading quality remains the same over

time. It is possible to extend the model for τ_{igt} to capture its change. We leave it for future investigation.

Joint Modelling of Student-Specific Latent Variables. The above model specification introduces four student-specific latent variables $(\alpha_i, \beta_i, \eta_i^2, \phi_i^2)$ for each student i . We allow for dependence between these latent variables, which will enable us to borrow information between the performance data and the grading data of the same student when evaluating their performances as an examinee and as a grader. More precisely, we assume $(\alpha_i, \beta_i, \eta_i^2, \phi_i^2)$, $i = 1, \dots, N$ are i.i.d., with $(\alpha_i, \beta_i, \log(\eta_i^2), \log(\phi_i^2))$ following a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_4)^\top$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{4 \times 4}$. For identifiability purposes, we fix $\mu_1 = \mu_2 = 0$, so that the average score of each assignment (averaged across students and graders) is completely captured by the difficulty parameter δ_i . We note that no constraint is imposed on μ_3 and μ_4 .

2.2 Bayesian Inference

We adopt a fully Bayesian procedure for drawing statistical inference under the proposed model.

Prior Specification. We first specify the prior for the assignment difficulty parameters $\delta_1, \dots, \delta_T$. When T is relatively small (e.g., $T \leq 5$), we simply assume each δ_i to have a weakly informative prior $N(0, 5^2)$. When T is larger, a hierarchical prior specification may be used by assuming $\delta_1, \dots, \delta_T$ to be i.i.d. following a certain prior distribution (e.g., normal) with some hyper-parameters and further setting a hyper-prior distribution for the hyper-parameters. We proceed to specify a prior for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the joint distribution for the student-specific latent variables. Recall that μ_1 and μ_2 are constrained to zero, for which no prior needs to be set. For μ_3 and μ_4 , we assume them to be independent, each following a weakly informative normal prior $N(0, 5^2)$. For the covariance matrix $\boldsymbol{\Sigma}$, we reparameterize

$$\boldsymbol{\Sigma} = \mathbf{S}\boldsymbol{\Omega}\mathbf{S},$$

where $\mathbf{S} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{44}})$ is a 4×4 diagonal matrix with its diagonal entries being the standard deviations of $(\alpha_i, \beta_i, \log(\eta_i^2), \log(\phi_i^2))$, and $\boldsymbol{\Omega} = (\omega_{ij})_{4 \times 4} = \mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}$ is the correlation matrix of $(\alpha_i, \beta_i, \log(\eta_i^2), \log(\phi_i^2))$. The prior distribution on $\boldsymbol{\Sigma}$ is imposed by the priors on $\mathbf{S}$ and $\boldsymbol{\Omega}$. For $\mathbf{S}$, we assume $\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{44}}$ to be i.i.d., each following a half-Cauchy distribution with location 0 and scale 5. For the correlation matrix $\boldsymbol{\Omega}$, we assume a Lewandowski-Kurowicka-Joe (LKJ) prior distribution with shape parameter 1 [6], which corresponds to the uniform distribution over the space of all correlation matrices.

Computational Aspects. We adopt the No-U-Turn Hamiltonian Monte Carlo (HMC) sampler [5], a computationally efficient MCMC sampler, and implement it under the Stan programming language. Compared with classical MCMC samplers, such as the Gibbs and Metropolis-Hastings samplers, the No-U-Turn HMC sampler uses geometric properties of the target distribution to propose posterior samples and thus converges faster to high-dimensional target distributions [5]. Regarding the implementation, for all

the models, 4 HMC chains are run in parallel for 2000 iterations, of which the first 1000 iterations were specified as warm-up. We use the `rstan` R package to analyze the resulting posterior samples, more specifically, it enables us to merge the MCMCs, compute the summary statistics of the posteriors and check the MCMC mixing and convergence. Moreover, the R package `loo` [8] and `Bayesplot` [3] are used separately for model comparisons and to plot the results, respectively.

3 Real Data Example

These peer grading data are from [10]. Participants are $N = 274$ American undergraduate students attending a Biology course. A double-blinded individual peer assessment was implemented for four different assignments, $T = 4$ throughout the course. Each student was graded, on average, by a random set of 5 other students for each assignment. The coursework was rated on a 1 – 7 Likert scale with instructor-provided anchor descriptions for each rating level. For the current analysis, only the students who completed at least three assignments were considered. This allows us to fit the LGC model. It results in a sample size of $N = 212$ students.

Results. No mixing or convergence issues emerge from a graphical inspection of the MCMCs and also as suggested by the values of the $\hat{R} < 1.01$.

The assignment difficulty levels seem to be in increasing order (see Table 1). Conditional to the other parameters, the first and the fourth assignments are, respectively, the easiest and the most difficult ones. The 95% quantile-based credible intervals of the assignment difficulty parameters are moderately narrow, suggesting a low level of uncertainty for these parameters. The estimates of the location parameters of the latent variables, μ_3 and μ_4 , suggest that students are on average more consistent than reliable. The posterior mean and the 95% credible intervals of the first quantity, respectively $\hat{\mu}_3 = -1.27$ and $(-1.46, -1.10)$, are considerably smaller than those related to the second one, $\hat{\mu}_4 = -0.46$ and $(-0.51, -0.41)$. This implies that, on average, the variance of the student's ability is smaller than the error variance of the grades they give. They are more consistent as an examinee than as a grader. From a substantive point of view, considering that they are not grader experts, it seems reasonable. Note that these quantities are expressed on a logarithmic scale, which implies that the average variance of the students' proficiency across different assignments is $\exp(\hat{\mu}_3) = 0.28$, and, on average, their reliability parameter is $\exp(\hat{\mu}_4) = 0.63$.

Students are moderately homogeneous in terms of their mean abilities, as suggested by the relatively small values of σ_1 , whereas they are more variable in their systematic bias, which is indicated by the posterior values of the parameter σ_2 . In other words, they are, on average, more similar as examinees than they are as graders. Moreover, students are widely different from each other concerning their consistency across assignments, as suggested by the posterior values of the structural parameter σ_3 . There is slightly less variability among them concerning the reliability parameters as suggested by the values of σ_4 .

Regarding the dependency among the latent variables, it emerges that higher values of students' proficiency are associated with higher values of consistency. Indeed, there is evidence of a strong correlation, between the first and the second student-specific

variable, respectively, α_i and $\log(\eta_i^2)$, as suggested by the posterior mean and the 95% credible interval $\hat{\omega}_{13} = -0.86$ and $(-0.96, -0.74)$, respectively. In addition, higher values of mean bias are predictive of higher reliability levels. This is evidenced by the posterior mean and 95% credible interval of the parameter ω_{24} , respectively, $\hat{\omega}_{24} = -0.73$ and $(-0.83, -0.63)$. The estimates of the other correlation parameters do not provide clear evidence about any other dependency among the latent variables under the present model.

Going to the student-specific level, each student might be provided with a Score estimate and a 95% quantile-based credible interval for each assignment as a measure of uncertainty. The posterior mean of $\hat{\theta}_{it} - \hat{\delta}_t$ might be a point estimate for students' Scores. The posterior distributions of both the average bias and the reliability of each grader might be useful information to assess their grading behaviour.

Table 1. Estimated structural parameters. For each parameter, the posterior mean (Post. Mean) and the 95% quantile-based credible interval (CI) are reported. The parameter δ_t is the difficulty level of the assignment t ; μ_3 and μ_4 are the location parameters of the third and the fourth latent variable; $\sigma_1, \dots, \sigma_4$ are the standard deviations of the latent variables; ω_{mn} is the correlation parameter between the latent variables m and n .

	Parameter	Post. Mean	95% CI
Assignments	δ_1	-6.31	$(-6.39, -6.24)$
	δ_2	-5.38	$(-5.46, -5.32)$
	δ_3	-5.36	$(-5.43, -5.29)$
	δ_4	-4.96	$(-5.03, -4.89)$
Students	μ_3	-1.27	$(-1.46, -1.10)$
	μ_4	-0.46	$(-0.51, -0.41)$
	σ_1	0.23	$(0.19, 0.28)$
	σ_2	0.35	$(0.32, 0.39)$
	σ_3	0.66	$(0.53, 0.83)$
	σ_4	0.32	$(0.29, 0.37)$
	ω_{12}	-0.09	$(-0.27, 0.08)$
	ω_{13}	-0.86	$(-0.96, -0.74)$
	ω_{14}	0.17	$(-0.02, 0.36)$
	ω_{23}	-0.08	$(-0.29, 0.13)$
	ω_{24}	-0.73	$(-0.83, -0.63)$
	ω_{34}	0.12	$(-0.10, 0.34)$

Acknowledgments. We are sincerely grateful to Professor Oscar Luaces for sharing the data set on peer grading.

References

1. Bartholomew, D., Knott, M., Moustaki, I.: *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd edn. Wiley, New York (2011)
2. Double, K.S., Mcgrane, J.A., Hopfenbeck, T.N.: The impact of peer assessment on academic performance: a meta-analysis of control group studies. *Educ. Psychol. Rev.* **32**, 481–509 (2020)
3. Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A.: Visualization in bayesian workflow. *J. R. Stat. Soc. A* **182**, 389–402 (2019)
4. Gamage, D., Staubitz, T., Whiting, M.: Peer assessment in MOOCs: systematic literature review. *Distance Educ.* **40**(2), 1–22 (2021)
5. Hoffman, M.D., Gelman, A.: The no-u-turn sampler. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014)
6. Lewandowski, D., Kurowicka, D., Joe, H.: Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**(9), 1989–2001 (2009)
7. Sanchez, C., Atkinson, K., Koenka, A., Moshontz, H., Cooper, H.: Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: a meta-analysis. *J. Educ. Psychol.* **109**(8), 1049–1066 (2017)
8. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017)
9. Yin, S., Chen, F., Chang, H.: Assessment as learning: how does peer assessment function in students' learning? *Front. Psychol.* **13**, 912568 (2022)
10. Zong, Z., Schunn, C.D., Wang, Y.: What aspects of online peer feedback robustly predict growth in students' task performance? *Comput. Hum. Behav.* **124**, 106924 (2021)



Machine Learning and Air Pollution: An Empirical Application for Northern Italy

Angelo Mimmo  and Daniele Toninelli[✉] 

Department of Economics, University of Bergamo, via dei Caniana, 2, 24127 Bergamo, Italy
daniele.toninelli@unibg.it

Abstract. Air pollution is a serious threat for almost all European countries. Recently in Northern Italy PM_{10} , $PM_{2.5}$ and NO_2 limits set by the EU were often overtaken. In order to fight this trend, it is crucial to analyse air pollutants. This work uses two machine learning techniques (Natural Cubic Splines and Bagging) to predict PM_{10} and NO_2 daily concentration for the years 2016 to 2019. We assess the methods accuracy by means of the Root Mean Square Error, the Weighted Mean Absolute Percentage Error and the Pearson's correlation between predictions and observed values. Bagging slightly outperforms Natural Cubic Splines and performances are better for NO_2 than for PM_{10} . We are able to provide estimates even when actual on field measurements are not available. Moreover, we could use these models to forecast pollutant levels for the next few days by plugging in estimates of our predictors. This approach provides national and local governments with tools to study the air pollution problem, shifting from ex-post measurements and sanctions to ex-ante strategies and corrections.

Keywords: Air pollution assessment · machine learning · Bagging · Natural Cubic Splines

1 Introduction

Air pollution is an extremely harmful threat for many European cities. In February 2024, in many cities of Northern Italy, the limits set by the EU for different pollutants were overtaken. The Copernicus Atmosphere Monitoring Service¹ analyzed the compounds of PM_{10} in Milan (one of the most affected cities), highlighting a prevalence of Primary Organic Matter, due to combustion (residential heating, transports, industries). Increasing $PM_{2.5}$ and NO_2 levels were also observed, confirming the human activities as primary source of pollution. Thus, it is crucial to analyse air pollutants levels and to provide tools to predict them.

In this work we apply two machine learnings techniques, Natural Cubic Splines (James et al., 2013) and Bagging (Breiman, 1996; James et al., 2013), to predict PM_{10} and NO_2 daily concentrations for the period 2016–2019. Predictors include weather, mobility, geographical and demographical variables. We analyse 34 cities of the North-West

¹ <https://atmosphere.copernicus.eu/extreme-episode-particulate-matter-air-pollution-across-italy-po-valley>; last accessed 2024/03/01.

and North-East NUTS 1 Italian regions. The performance of these models is assessed by means of Root Mean Square Error (RMSE), Weighted Mean Absolute Percentage Error (WMAPE) and Pearson's correlation between predicted and actual values. Our results show that both models perform well, although Bagging slightly outperforms Natural Cubic Splines. Moreover, performances are better for NO_2 than for PM_{10} .

Our scope is to assess air pollutants levels using machine learning algorithms, providing estimates even when on field measurements are not available. These models could be also used to forecast pollutant levels for the next two/three days by plugging in estimates of our predictors. Thus, these tools can help national and local governments (regions, provinces, municipalities) to study and to address the air pollution problem.

2 Literature Review

Air pollution is defined as the qualitative or quantitative variation of the standard composition of the atmosphere in a specific geographical area (Campanile et al., 2021). Pollutants are divided in two categories (Daly and Zannetti, 2007): 1) Primary pollutants directly emitted into the atmosphere (e.g., carbon, nitrogen, sulphur and halogen compounds, particulate matters); 2) Secondary pollutants produced from primary pollutants (e.g., nitrogen dioxide and ozone). Recently, machine learning techniques have been increasingly applied to environmental data, in particular for studying air pollutants. For example, Campanile et al. (2021) applied several machine learning methods to forecast hourly carbon monoxide (CO) levels in Naples. Ridge Regression (Hoerl and Kennard, 1970) was found as the best method in terms of R^2 , Mean Absolute Error and Mean Squared Error. Stafoggia et al. (2019) estimated daily PM_{10} , $\text{PM}_{2.5}$ and $\text{PM}_{2.5-10}$ concentrations for Italy (using 1-km² grid resolution) applying Random Forest (first proposed by Breiman, 2001). Several predictors were considered (e.g., Aerosol Optical Depth, weather variables and Normalized Difference Vegetation Index). However, no other pollutant was included as predictor. The results showed a very good predictive power of Random Forest, with a R^2 close to 0.95 and negligible RMSE (~ 0.02).

Compared to the previous literature, in this paper we will consider a more local perspective: Northern Italy. We estimate PM_{10} and NO_2 daily concentrations using several weather, mobility, geographical and demographical variables. We apply two different machine learning methods: Natural Cubic Splines (James et al., 2013) and Bagging (Breiman, 1996; James et al., 2013). With these tools, local governments can implement *ad hoc* actions on the affected areas before pollutants concentrations reach a peak.

3 Data Collection and Methodology

This paper focuses on the 34 main Italian cities of the North-West and North-East NUTS 1 Italian regions for which pollutants data from monitoring station were available (Fig. 1). Our data include daily air pollutants (PM_{10} , NO_2 and other compounds) for 66 Italian cities, for years 2016 to 2019. We focus on the Po Valley area, since it is more severely affected by air pollution. Data were downloaded from the European Environment Agency (EEA) website. They are collected by the Italian Institute for Environmental Protection

and Research (ISPRA). We relied on primary validated assessment (E1a) data, verified by ISPRA. Daily pollutants concentrations were averaged over all the available stations, resulting in a single average value for each city and for each day. Daily weather predictors were downloaded from the NASA Prediction Of Worldwide Energy Resources website². These are not measured on-field but derived from satellite observations. Raw satellite data are stored, analyzed and modelled by the NASA Global Modelling and Assimilation Office. Geographical and demographical variables and proxies for mobility and traffic were collected from the Italian National Institute of Statistics (ISTAT). Table 1 lists all the variables used.



Fig. 1. Norther Italy cities included in the analysis.

The two machine learning models used in this paper are different from the methodological point of view. **Natural Splines** belong to the so-called parametric methods. They are a subcategory of Regression Splines, which, in turn, are a subclass of Piecewise Polynomials. These latter involve fitting separate low-degree polynomials over different regions of a set of predictors X , instead of a high-degree polynomial over its entire range (James et al., 2013). This way, we estimate region-specific coefficients that change in correspondence of the K knots. However, Piecewise Polynomials tend to be discontinuous in correspondence of these knots. Regression Splines require the fitted curve and its first and second derivatives to be continuous at each knot. This way, the resulting piecewise polynomial functions are continuous and smooth in correspondence of the knots. Nevertheless, Regression Splines can suffer “from high variance at the outer range of the predictors” (James et al., 2013, p. 274). Natural Splines fix this by being linear in the regions where X is smaller than the smallest knot and larger than the largest knot, producing more stable estimates at the boundaries. We rely on a Natural Cubic Spline ($d = 3$). The knots are uniformly placed over the range of X ; their number is defined using a validation set and considering degrees of freedom between 1 and 20.

² NASA POWER website: <https://power.larc.nasa.gov>, last accessed 2024/03/03.

Table 1. Target variables and predictors used.

Variable type	Variable	Unit of measurement	Source
Air pollutants	Nitrogen dioxide (NO ₂)	µg/m ³	EEA
	Coarse particulate matter (PM ₁₀)		
Weather	Precipitations	mm	NASA
	Precipitations in the previous day	mm	
	Relative humidity	%	
	Atmospheric pressure	kPa	
	Temperature	°C	
	Wind speed	m/s	
	All sky insolation	kWh/m ²	
Geographical, demographical and mobility	Latitude and longitude	°	ISTAT
	Average altitude	m	
	Yearly population density	Residents/km ² of town surface	
	Yearly total green space density	% of town surface	
	Yearly provincial motorization rate	Vehicles/1000 residents	
	Yearly provincial vehicle density	Vehicles/ km ² of urbanized surface	
	Yearly provincial polluting potential index	High-medium polluting/100 med.-low polluting potential vehicles	

Conversely, **Bootstrap Aggregation (Bagging)** is a non-parametric method suited for large, high dimensional datasets “where finding a good model or classifier in one step is impossible because of the complexity and scale of the problem” (Bühlmann and Yu, 2002, p. 927). It is an ensemble averaging method that takes predictions from several Regression Trees and then averages them (Breiman, 1996). B samples are selected with replacement from the training set and a Regression Tree is fit on them. Bagging relies on the fact that the mean $\bar{X}$ of n i.i.d. random variables (each with variance σ^2) has a variance equal to $\sigma^2/n < \sigma^2$. B is not critical: a very large value will not cause overfitting (James et al., 2013). The main issue of B is computational. Indeed, the final model has been trained using $B = 50$, due to the limited capacity of the local computing environment where both methods have been deployed. The full dataset has been split into three sets: 1) Training set (60% of observations), used to fit the models. 2) Validation set (20%), used to select the best value for the hyperparameters which control the Bias-Variance

trade-off. 3) Test set (20%), used to assess the performance of the models, with the hyperparameters best values, on data not previously used to train them. To perform this split, we implemented a stratified random sampling, using two auxiliary variables: the observation month and the group to which the city belongs to. The groups have been defined applying agglomerative hierarchical clustering on the whole set of 66 cities. The inputs for this Cluster Analysis are the 16 principal components obtained applying the PCA on both yearly means of the predictors and their differences. This led to 4 clusters of similar cities and to 48 strata. We used the validation set to select the best degrees of freedom value (i.e., the number of knots) for the Natural Cubic Spline: $DF = 20$. We retrained the model on the aggregation of the training and validation set (80% of observations). Not being B a critical parameter, we directly fitted Bagging on this full training set, with $B = 50$. We assessed the models performance by using the Root Mean Square Error (RMSE), the Pearson's correlation coefficient between predicted and actual values and the Weighted Mean Absolute Percentage Error (WMAPE).

4 Results

Models show a good predictive power for both pollutants (Table 2): the Pearson correlation coefficient between predicted and actual values is above 0.7 and the WMAPE is below 35%. For PM_{10} shifting from Natural Cubic Spline to Bagging leads to a gain of 7.32% in the WMAPE (33.71% to 26.39%). For NO_2 this gain is smaller, around 3.27% (20.92% to 17.65%). These findings are confirmed by the other two performance metrics. The predictive superiority of ensemble methods as Bagging is widely documented in the literature. In Stafoggia et al. (2019), for example, Random Forest (a slight variation of Bagging) showed very good predictive performances for PM_{10} , $PM_{2.5}$ and $PM_{2.5-10}$. Another interesting result is that, regardless of the model considered, the predictive performance is higher for NO_2 than for PM_{10} . This means that trying to predict PM_{10} by using meteorological and mobility variables is a more difficult task, if compared to NO_2 . According to Cameletti (2020), the formation of PM_{10} is complex, with sources of primary particulate matter (e.g. household heating, road traffic, industrial combustion, agriculture) as well as chemical reactions with other pollutants. For example, in the Po Valley primary emissions are only the second cause of particulate matter in wintertime. The main source is represented by Secondary Organic Aerosol (SOA) generated by chemical reactions in the atmosphere (Larsen et al., 2012).

Table 2. Machine learning models performance on test set.

Pollutant	ML model	RMSE	Correlation	WMAPE %
NO ₂	Natural Cubic Spline ($DF = 20$)	8.87	0.81	20.92%
	Bagging ($B = 50$)	7.54	0.87	17.65%
PM ₁₀	Natural Cubic Spline ($DF = 20$)	13.46	0.70	33.71%
	Bagging ($B = 50$)	10.79	0.82	26.39%

5 Conclusions

Both Natural Cubic Spline and Bagging led to a good predictive performance, although the latter seems to be better suited for estimating air pollutants concentrations. In addition, in line with the literature on this topic, predicting PM_{10} appears to be more difficult than for NO_2 . In any case, the overall good results show how it is possible to develop and provide predictive tools, similar to the ones presented in this paper, to national and local governments. This way, the policies and the actions to tackle air pollution could be designed not only on a local basis, but also implemented before this problem reaches its peak. For example, the block of most polluting vehicles in urban areas could be enforced before a certain pollutant overcomes the legally binding limit set by the EU. Similarly, remote working could be incentivized, when the predicted pollutants levels show a potential spike. Our local focus should also provide a more detailed picture of the pollution issue, encouraging to coordinate more efficiently local and general policies. Moreover, our findings could contribute to reverse the current air pollution regulation: from *ex-post* measures and sanctions to *ex-ante* initiatives and corrections.

A possible further development of this analysis with more recent data could help in assessing whether our findings are confirmed for a different and more recent time frame and/or on a wider area. Moreover, additional machine learning methods could be experimented (e.g., other ensemble methods as Random Forest or Gradient Boosting). This is expected to have potential beneficial effects on the predictive performance for both NO_2 and PM_{10} . Finally, the inclusion of additional predictors frequently used in the literature (as Aerosol Optical Depth) could improve the accuracy of estimates. This holds, in particular, for PM_{10} , given that trying to predict it by using the current set of meteorological and mobility variables is a more difficult task, if compared to NO_2 .

Funding. Project supported by 60% University funds (University of Bergamo, 2023 and 2024).

References

- Breiman, L.: Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996)
- Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
- Bühlmann, P., Yu, B.: Analyzing bagging. *Ann. Stat.* **30**(4), 927–961 (2002)
- Cameletti, M.: The effect of Corona Virus lockdown on air pollution: evidence from the city of Brescia in Lombardia Region (Italy). *Atmos. Env.* **239**, 1–6 (2020)
- Campanile, M., Cantiello, P., Iacono, M., Lotito, R., Marulli, F., Mastroianni, M.: Applying machine learning to weather and pollution data analysis for a better management of local areas: the case of Napoli, Italy. In: *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security*. SciTePress (2021)
- Daly, A., Zannetti, P.: An introduction to air pollution – definitions, classifications, and history. In: Zannetti, P., Al-Ajmi, D., Al-Rashied, S. (eds.) *Ambient air pollution*, pp. 1–14. Fremont (2007)
- Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
- James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning*. Springer, New York (2013)

- Larsen, B., Gilardoni, S., Stenström, K., Niedzialek, J., Jimenez, J., Belis, C.: Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmos. Env.* **50**, 203–213 (2012)
- Stafoggia, M., et al.: Estimation of daily PM₁₀ and PM_{2.5} concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Env. Int.* **124**, 170–179 (2019)



Big Data Subsampling: A Review

Rosa Arboretti¹, Marta Disegna², and Alberto Molena²(✉)

¹ Department of Civil Environmental and Architectural Engineering,
University of Padova, Via Marzolo, 9, 35131 Padova, Italy

rosa.arboretti@unipd.it

² Department of Management and Engineering, University of Padova,
Stradella San Nicola, 3, 36100 Vicenza, Italy

marta.disegna@unipd.it, alberto.molena.1@phd.unipd.it

Abstract. Big data subsampling encompasses a set of techniques aimed at selecting the most informative subset from an initial dataset, with the purpose of conducting inferential analyses on said subset, thereby reducing the computational power required to perform such analyses. We present a literature review on the topic with an orientation specifically designed for practitioners. As such we will introduce some metrics used to analyze the sample of articles in the review and discuss them, then briefly present one of the most commonly used subsampling methods, highlighting its advantages and limitations.

Keywords: Big data · Subsampling · Review · Practitioners

1 Introduction

Big data have historically been characterized by three fundamental dimensions: volume, variety, and velocity [1]. Over the past few years, there has been a notable escalation in the volume of data generated, as observed in everyday experiences. Moreover, instances where data generation reaches truly massive scales are becoming increasingly prevalent. For instance, the Square Kilometre Array, recognized as the largest radio telescope in the world, is estimated to produce approximately 600 petabytes of data annually, equating to an average of over 1500 gigabytes of data on a daily basis [2]. Clearly, the fact that datasets of such dimensions are produced brings with them some significant challenges: for example, the computational power required to analyze these data is substantial and most likely prohibitive. Considering then more generally the fact that the amount of data being produced is increasing exponentially, the need has arisen for methodologies that allow the analysis of datasets of such large dimensions, with the computational power available. In the literature, three relevant approaches are present: divide and conquer, online updating, and subdata selection [3]. The latter approach consists of obtaining a sample from the total dataset that is as informative and effective as possible, in order to carry out the statistical inference phase using the sample, with the possibility of extending the

results to the entire dataset. Clearly, the focal point of this approach, which is also one of the topic's key challenges, is the method used to create these samples. In this contribution, our aim is to present a literature review on subdata selection methodologies, addressed primarily at practitioners. We will then briefly present, in Sect. 2, the parameters used to conduct the review and in Sect. 3 we will present a discussion on the main findings of the literature review. In Sect. 4 we will present one of the most cited methods, presenting its main advantages and limitations of application, thus providing a guide to practitioners. Lastly, Sect. 5 will encapsulate the key insights of this paper, thereby concluding our contribution.

2 Methodology

We undertook a literature review in December 2023, exploring the Scopus database. A query was formulated encompassing three distinct fields: Optimality, Big Data, and Selection. The various keywords utilized within these fields were sought within the title, abstract, and keywords sections. Subsequently, several filters were applied, primarily concerning the domain, where only papers falling under Computer Science, Engineering, Business, Management and Accounting, Mathematics, and Decision Science were retained. Additional filters were used pertaining to language, with only English-language papers being retained, and document type, where only articles and reviews were considered. The next phase involved the careful reading of abstracts to ascertain whether or not the retained documents aligned with the thematic focus. This process yielded 48 papers, each of which underwent comprehensive analysis following a predetermined framework.

3 Main Findings

Initially, we analyzed the temporal trend of the 48 examined publications and found that 34 papers [4–37] (almost 71% of the total) were published from 2021 onwards, indicating a growing interest in this research topic in recent years. As previously pointed out, the primary target audience of this review is practitioners. Consequently, the analytical framework employed primarily encompassed four dimensions: research methodology, to ascertain if studies apply these techniques to case studies; the mathematical criteria used to obtain the subsample; which machine learning models were used alongside these techniques; and limitations of the presented techniques. From analysis of the research methodology, it emerged that the vast majority of papers, 41 out of 48, employed both simulation studies and real case studies, indicating an approach oriented not only towards theory but also towards practice. Regarding the criteria used to obtain the subsample, analysis revealed that the most commonly utilized criteria were A-optimal, used in 18 out of 48 papers [5, 8, 11, 12, 16, 17, 23, 26, 29, 32, 36, 38–44], L-optimal, used in 16 papers [5, 8, 12–14, 16, 17, 22, 23, 25–27, 32, 34–36, 40–44], and D-optimal, used in 7 papers [7, 11, 15, 20, 35, 45, 46], derived from the

design of experiments. Briefly, these three criteria can be outlined as follows: A-optimal aims to minimize the trace of the inverse of the information matrix, ensuring a low correlation between estimated coefficients of statistical models; L-optimal focuses on minimizing the determinant of the information matrix, maximizing the precision of parameter estimates in the model; D-optimal aims to minimize the maximum variance of individual coefficients in the model, ensuring an efficient and accurate estimation of parameters. Studies employing space-filling design techniques, however, appear to be relatively limited [6,9,24]. Let us now analyze the ML models used after the subsampling phase; in doing so, it is necessary to specify that many subsampling methods depend directly on the ML model we intend to apply. In fact, as seen previously in the explanation of the optimality criteria, the goal is generally to reduce the variation in parameter estimation when a certain model is used. This implies that the method to obtain the subdata is optimal only when used in combination with the model for which it was constructed and therefore loses this optimality characteristic when used in combination with other models. This creates a problem, as discussed in [9], namely that from a statistical point of view, we would like the relationship between predictors and response variable to be correctly approximated without assuming that it is satisfied by a certain model chosen upstream of the data analysis process. Nevertheless, in this review, we aim to analyze the existing literature, and from the analysis performed, we noticed that the majority of existing methodologies are related to the linear regression model; indeed, in 10 out of 48 articles this model was used [7,13,14,20,23,29,33,37,46,47]. The second most frequent ML model was the logistic regression model, which was used in 8 papers [18,23,28,29,32,42,44,45]. There were also a significant number of applications for quantile regression, with 4 papers [5,34,40,48], and generalized linear models, with 3 papers [17,36,41]. Finally, from the previous discussion regarding the relationship between ML models and subsampling methodologies, a significant limitation in the field, confirmed by our analysis, is the need to choose the model to apply a priori in order to apply criteria that create optimal datasets. Indeed, from all the analyzed papers, we found only 4 contributions [6,9,31,35] presenting methodologies that can be defined as model-agnostic, i.e. independent of the model, and two of those use space-filling design methodologies to select the subset [6,9]. It is therefore possible to infer that the idea of using space-filling design criteria to obtain subsets of data that are as similar as possible to the initial dataset is an idea that is beginning to be developed; indeed, these articles were published in 2022 and 2023, and could prove to be successful in overcoming this limitation.

4 IBOSS: Advantages and Limitations

In this section we briefly summarize one of the most cited methods widely referenced and frequently employed as a benchmark in various comparative tests. The method should be used in conjunction with the linear regression model and is presented in [46]. Said method is IBOSS (Information-Based Optimal

Subdata Selection). The idea underlying this methodology is the deterministic selection of points that are most informative, so that the subset retains much of the information contained in the initial dataset. This methodology is based on the D-optimality criterion and tends to select points at the boundary of the design space. In general, it offers three significant advantages: it is faster than many existing methods, can be used with distributed parallel computation, and the variance of parameter estimators converges to 0 even if the subdata size is fixed. However, this methodology has some limitations. Firstly, as discussed in the previous section, it returns an optimal subsample only when we consider data underlying a linear model; if the relationship is different, this is not true [31]. Additionally, as mentioned earlier, the method tends to select points at the extremes of the boundaries, but this can be a strong limitation when outliers are present, as pointed out in [7].

5 Conclusion

In this short paper we have presented a literature review conducted on big data subsampling. From the presented analysis it emerges that many methods examined in the review are based on traditional optimality criteria of design of experiments. Indeed, we demonstrated that the A-optimal, D-optimal, and L-optimal criteria are the three most employed to create the subsample. We also discussed the models with which these approaches have been employed and briefly outlined the advantages and disadvantages of IBOSS, one of the most cited methods. Our analysis also revealed that an intriguing area for development appears to be associated with the utilization of space-filling design techniques, aimed at creating methodologies that are ML-agnostic.

References

1. Laney, D., et al.: 3D data management: controlling data volume, velocity and variety. META Group Res. Note **6**(70), 1 (2001)
2. Thavasimani, P.: Rapidxfer-data transfer framework for square kilometre array. J. Data Inf. Manag. **3**, 251–260 (2021)
3. Wang, C., Chen, M.H., Schifano, E., Wu, J., Yan, J.: Statistical methods and computing for big data. Stat. Interface **9**(4), 399 (2016)
4. Jiang, Z., Wu, Y., Wang, M., Wu, L.: A novel residual subsampling method for skew-normal mode regression model with massive data. Commun. Stat. Theory Methods **53**, 1–17 (2023)
5. Shao, L., Song, S., Zhou, Y.: Optimal subsampling for large-sample quantile regression with massive data. Can. J. Stat. **51**(2), 420–443 (2023)
6. Yi, S.Y., Zhou, Y.D.: Model-free global likelihood subsampling for massive data. Stat. Comput. **33**(1), 9 (2023)
7. Deldossi, L., Pesce, E., Tommasi, C.: Accounting for outliers in optimal subsampling methods. Stat. Pap. **64**, 1–17 (2023)
8. Yan, Q., Li, H., Niu, C.: Optimal subsampling for functional quantile regression. Stat. Pap. **64**(6), 1943–1968 (2023)

9. Deng, L.Y., Yang, C.C., Bowman, D., Lin, D.K., Lu, H.H.S.: Big data model building using dimension reduction and sample selection. *J. Comput. Graph. Stat.* **33**, 1–13 (2023)
10. Ren, M., Zhao, S., Wang, M., Zhu, X.: Robust optimal subsampling based on weighted asymmetric least squares. *Stat. Pap.* **65**, 1–31 (2023)
11. Yu, J., Liu, J., Wang, H.: Information-based optimal subdata selection for non-linear models. *Stat. Pap.* **64**, 1–25 (2023)
12. Keret, N., Gorfine, M.: Analyzing big EHR data-optimal cox regression subsampling procedure with rare events. *J. Am. Stat. Assoc.* **118**, 1–14 (2023)
13. Ren, M., Zhao, S., Wang, M.: Optimal subsampling for least absolute relative error estimators with massive data. *J. Complex.* **74**, 101694 (2023)
14. Zhang, H., Zuo, L., Wang, H., Sun, L.: Approximating partial likelihood estimators via optimal subsampling. *J. Comput. Graph. Stat.* **33**, 1–27 (2023, just-accepted)
15. Reuter, T., Schwabe, R.: Optimal subsampling design for polynomial regression in one covariate. *Stat. Pap.* **64**, 1–23 (2023)
16. Chao, Y., Huang, L., Ma, X., Sun, J.: Optimal subsampling for modal regression in massive data. *Metrika* **87**, 1–31 (2023)
17. Mahendran, A., Thompson, H., McGree, J.M.: A model robust subsampling approach for generalised linear models in big data settings. *Stat. Papers* **64**, 1–21 (2023)
18. Kirkby, J.L., Nguyen, D.H., Nguyen, D., Nguyen, N.N.: Inversion-free subsampling newton’s method for large sample logistic regression. *Stat. Papers* **63**, 1–21 (2022)
19. Chang, M.C.: Predictive subdata selection for computer models. *J. Comput. Graph. Stat.* **32**(2), 613–630 (2023)
20. Yu, J., Wang, H.: Subdata selection algorithm for linear model discrimination. *Stat. Papers* **63**(6), 1883–1906 (2022)
21. Teng, G., Tian, B., Zhang, Y., Fu, S.: Asymptotics of subsampling for generalized linear regression models under unbounded design. *Entropy* **25**(1), 84 (2022)
22. Wang, J., Zou, J., Wang, H.: Sampling with replacement vs poisson sampling: a comparative study in optimal subsampling. *IEEE Trans. Inf. Theory* **68**(10), 6605–6630 (2022)
23. Wang, J., Wang, H., Xiong, S.: Unweighted estimation based on optimal sample under measurement constraints. *Can. J. Stat.* **52**, 291–309 (2022)
24. He, L., Hung, Y.: Gaussian process prediction using design-based subsampling. *Stat. Sin.* **32**(2), 1165–1186 (2022)
25. Wang, T., Zhang, H.: Optimal subsampling for multiplicative regression with massive data. *Stat. Neerl.* **76**(4), 418–449 (2022)
26. Shao, Y., Wang, L.: Optimal subsampling for composite quantile regression model in massive data. *Stat. Pap.* **63**(4), 1139–1161 (2022)
27. Yuan, X., Li, Y., Dong, X., Liu, T.: Optimal subsampling for composite quantile regression in big data. *Stat. Pap.* **63**(5), 1649–1676 (2022)
28. Song, Y., Dai, W.: Deterministic subsampling for logistic regression with massive data. *Comput. Stat.* **39**, 1–24 (2022)
29. Lee, J., Schifano, E.D., Wang, H.: Fast optimal subsampling probability approximation for generalized linear models. *Econ. Stat.* **29**, 224–237 (2021)
30. Wang, H., Kim, J.K.: Maximum sampled conditional likelihood for informative subsampling. *J. Mach. Learn. Res.* **23**(1), 1–50 (2022)
31. Joseph, V.R., Mak, S.: Supervised compression of big data. *Stat. Anal. Data Mining ASA Data Sci. J.* **14**(3), 217–229 (2021)
32. Zuo, L., Zhang, H., Wang, H., Sun, L.: Optimal subsample selection for massive logistic regression with distributed data. *Comput. Stat.* **36**, 1–28 (2021)

33. Wang, L., Elmstedt, J., Wong, W.K., Xu, H.: Orthogonal subsampling for big data linear regression. *Ann. Appl. Stat.* **15**(3), 1273–1290 (2021)
34. Wang, H., Ma, Y.: Optimal subsampling for quantile regression in big data. *Biometrika* **108**(1), 99–112 (2021)
35. Deldossi, L., Tommasi, C.: Optimal design subsampling from big datasets. *J. Qual. Technol.* **54**(1), 93–101 (2021)
36. Ai, M., Yu, J., Zhang, H., Wang, H.: Optimal subsampling algorithms for big data regressions. *Stat. Sin.* **31**(2), 749–772 (2021)
37. Ren, M., Zhao, S.L.: Subdata selection based on orthogonal array for big data. *Commun. Stat. Theory Methods* **52**(15), 5483–5501 (2023)
38. Yu, J., Wang, H., Ai, M., Zhang, H.: Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *J. Am. Stat. Assoc.* **117**(537), 265–276 (2022)
39. Zhang, H., Wang, H.: Distributed subdata selection for big data via sampling-based approach. *Comput. Stat. Data Anal.* **153**, 107072 (2021)
40. Ai, M., Wang, F., Yu, J., Zhang, H.: Optimal subsampling for large-scale quantile regression. *J. Complex.* **62**, 101512 (2021)
41. Gao, J., Wang, L., Lian, H.: Optimal decorrelated score subsampling for generalized linear models with massive data. *Sci China Math* **67**, 1–26 (2023)
42. Wang, H.: More efficient estimation for logistic regression with optimal subsamples. *J. Mach. Learn. Res.* **20**, 1–59 (2019)
43. Yao, Y., Wang, H.: Optimal subsampling for softmax regression. *Stat. Pap.* **60**, 585–599 (2019)
44. Wang, H., Zhu, R., Ma, P.: Optimal subsampling for large sample logistic regression. *J. Am. Stat. Assoc.* **113**(522), 829–844 (2018)
45. Cheng, Q., Wang, H., Yang, M.: Information-based optimal subdata selection for big data logistic regression. *J. Stat. Plan. Inference* **209**, 112–122 (2020)
46. Wang, H., Yang, M., Stufken, J.: Information-based optimal subdata selection for big data linear regression. *J. Am. Stat. Assoc.* **114**(525), 393–405 (2019)
47. Ma, P., Sun, X.: Leveraging for big data regression. *Wiley Interdiscip. Rev. Comput. Stat.* **7**(1), 70–76 (2015)
48. Fan, Y., Liu, Y., Zhu, L.: Optimal subsampling for linear quantile regression models. *Can. J. Stat.* **49**(4), 1039–1057 (2021)



Survey on Urban Environmental Data: Data Validation

Domenico Adamo¹, Gianpiero Bianchi², and Lucia Mongelli²(✉)

¹ Istat, Dipartimento per la Produzione Statistica, Direzione Centrale per le Statistiche Ambientali e Territoriali, Roma, Italy

adamo@istat.it

² Istat, Dipartimento per la Produzione Statistica, Direzione Centrale per la Raccolta Dati, Roma, Italy

{gianbia,mongelli}@istat.it

Abstract. The National Statistical Programme in force includes the “Survey on urban environmental data”, carried out annually by Istat in the provincial and metropolitan capitals. The process provides a comprehensive picture of the quality of the urban environment, the state, pressure indicators and monitors local government policies to improve the quality of the environment in cities. The whole process handles about 500 elementary variables. The work describes the study and design of a new validation process, from a general perspective, which includes automatic procedures for checking the consistency of the collected data, the storage of data and the interaction with the Municipal Statistical Offices. The proposed methodological approach uses a metalanguage to represent rules in formal logic. This methodology makes it possible to build a highly automated and generalized validation process. The result is greater process efficiency, higher data quality and reduced resource consumption. Automatic interoperability with the municipalities helps to increase the efficiency of the overall process. The system reduces the monitoring load by automating most tasks such as the interception of measurement errors, historical series discontinuity and other abnormal values.

Keywords: Urban environmental · re-engineering · official statistics · automation · Propositional Logic

1 Overview

The “Survey on urban environmental data”, included in the National Statistical Programme currently in force and carried out annually by Istat, collects environmental information relating to 110 Italian cities (95 provincial capitals, 14 metropolitan capitals plus the municipality of Cesena, which participates on a voluntary basis). The data and statistical information have the objective to supply an informative picture to support of the monitoring of the state of the urban environment and of the activities put in existence from the local administrations in order to assure the good quality of the atmosphere in the cities. The data on the urban environment is a multi-source statistical process, which

produces environmental indicators and is organized into 8 thematic modules: air quality, urban waste, mobility, noise, energy, urban green, water, eco management [1]. The survey produces and updates every year about 170 indicators, which make up a training framework as rich as heterogeneous, both for the variety and breadth of the themes, and for the complexity of the information collected and processed (see Fig. 1).

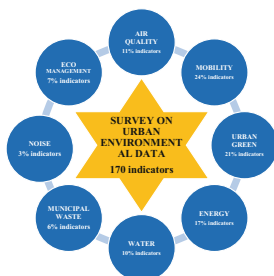


Fig. 1. Diagram on theme composition of the Survey on urban environmental data

All indicators are disseminated by municipality, and aggregate estimates are provided by geographical area. The survey indicators are also used to monitor some significant measures for the “Statistical Information for the 2030 Agenda in Italy”, in particular for the Goal 11 “Make cities and human settlements inclusive, safe, resilient and sustainable” [2].

The indicators produced by the survey and its ancillary sources span the entire spectrum of the DPSIR classification, developed by the European Environment Agency (EEA). This subdivides the environmental indicators in five categories (*Driving forces*, *Pressures*, *States*, *Impacts*, *Responses*), referable to as many phases of a general model of the interaction between man and environment (see Fig. 2).

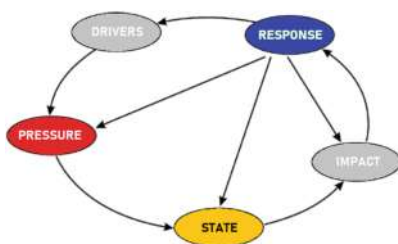


Fig. 2. The DPSIR model

The DPSIR model constitutes a reference structure, which allows to place the environmental data, relating to particular aspects of the observed phenomena, in the context of the chain of causes and effects that characterizes any environmental phenomenon, including the policies implemented towards it.

The indicators used to de-write each phase of the model are: determinants describing the driving forces, that is, the social, demographic and economic dynamics and the

corresponding changes in lifestyles, needs and levels of consumption and production, that generate pressure on the environment; *pressure* that measures the pressures on the environment exerted by driving forces (exploitation of resources, pollution emissions, waste production, etc.), which alter the state of the environment; *state*: they measure the qualities and characteristics of the environment and of the environmental resources, from which alterations can derive significant impacts on health, availability of resources, Biodiversity, etc., impacts that measure the effects of changes in the state of the environment in terms of environmental, economic, health, etc.; *response* describing the policies implemented to cope with pressures or mitigate impacts, which may take the form of objectives, plans or programmes, specific measures, etc. Particular attention shall be paid to the analysis of response indicators, defining the behavior of municipalities in the field of the environment and the use of the plan programming tools to improve the transparency of administrative processes and the active participation of citizens [7].

2 Data Collection

The data are collected by the Municipal Statistical Offices, which identify in the Administrations to which they belong the referents of the survey issues. Since 2018 there has been an integrative method of data collection, conducted before the survey, together with the update of the list of respondents, which uses the application LimeSurvey, distributed under the GNU GPL for the acquisition of georeferenced data (shapefile) on green areas (Urban Green theme) and on temporary events with authorisations derogating from noise limits (Noise issue).

According to the provisions of the Code of Digital Administration in 2005 (d.lgs 82/2005 and subsequent additions and changes) which provides that data must be transmitted to Istat in computerised mode, the CAWI technique (Computer Assisted Web Interviewing) has been introduced for data acquisition in electronic format, through the Gino++ (Gathering information Online) portal of Istat [9].

GINO++ is a generalized software that allows not only the collection of data but the complete management of surveys via web, creation of web questionnaire, controlled acquisition of data online and/or file upload, custom site preparation for the survey, monitoring of the status of questionnaires and records, contacts for reminders and reminders, reports. The questionnaires incorporate automatic control rules that signal inconsistent or abnormal data and non-response to respondents and monitors and, in the presence of critical breaches, prevent questionnaire submission.

3 Process Innovations

In line with the objectives of Istat to provide the country with correct statistical information and to innovate the various processes of production of statistical information, consistent with the progressive digitalization of data collection processes, it became necessary to design the use of innovative solutions by re-engineering the validation phase of the questionnaires, with the implementation of additional automatic control rules different from those already provided for by the validity and internal consistency checks of the Gino++ system.

The new rules manage, at least in part, the aspects so far entrusted to the review by monitors: in particular the interception of measurement errors, discontinuity of the time series and other anomalous values. These rules would work on a dynamic basis, by comparing the data collected in the current edition with those validated and disseminated in previous editions. In Table 1 the phases of the investigation after the re-engineering [3].

Table 1. Survey stages. After re-engineering

Questionnaire states	Municipal Offices	Stages	ISTAT		
			Data collection		Environmental statistics
			Survey design	Survey organization, Implementa- tion of CAWI questionnaires	Information contents and metadata management
			Data collection	Controlled acquisition through Gino electronic questionnaire with rules Automatic control of: -) measurement errors, -) historical series discontinuities -) other abnormal values. Monitoring of survey operations	Assistance to respond- ents and to data collec- tion staff
				Automated return on respond- ents	
		Data processing	Data editing and valida- tion		
		Data disse- mination	Data analysis and report- ing		

4 A Generalized Data Editing

A generalized editing system verifies the collected data consistency with respect to the survey by intra-record and inter-record check rules. The editing system also identifies inconsistency and redundancy in the rule set itself. The application catalogues the exact and wrong questionnaires, marking the collected units involved in the violations and specifying the fields involved in the violation of the rules [6]. The application uses a customizable metadata table to apply the control plan [5]. Checked output is properly logged out. Metadata table contains the following information:

- Editing Rules are classified by the following item types depending on their own formulation: Validity, Logic, Mathematics and Logical-Mathematics;
- The textual description of the rule;
- Rules are represented in machine readable formal language based on propositional logic;
- Editing Rules are also classified by the possibility to block the validation process until further correction. Items used are named “hard” (blocking rule) and “soft” (non-blocking rule);
- A hierarchy of rules describes the dependence and order of control of the rules. When a rule generates a violation, all the subsequent related rules can be directly marked as violated.

This section introduces the representation of rules in formal logic. In particular, it provides the definition and representation of check rule set and how to transform the textual rules in compatibility (or incompatibility rules) expressed in a formal language [4]. Rules are expressions typically used to detect and verify conditions occurring in a possibly large set of elements. In order to verify a set of checking rules, it is convenient to express them using a structure based on propositional logic [8].

Propositional logic, sometimes called sentential logic, builds over a grammar made of atomic building blocks and logical connectives to formalize complex statements. These atomic propositions that are unanalyzed because are known for the specific domain of application. A truth assignment to such atomic propositions will determine the “truth value” of the whole formula according to the truth rules of the logical connectives.

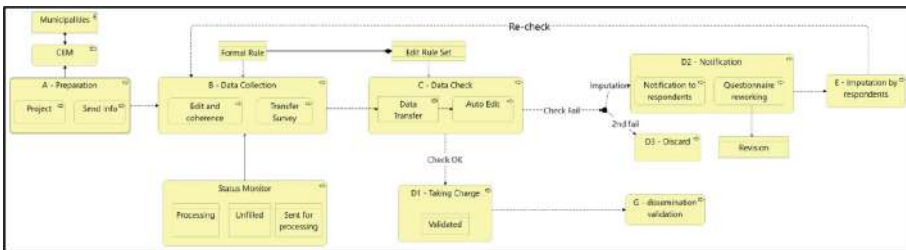
The traditional (symbolic) approach to propositional logic divides the syntactical and semantical functions:

- Syntax defines the structure of logical formulas and the relationship between its atomic propositions.
- Semantics, on the other hand, gives the interpretation and meaning associated with the syntactical objects.

Propositional calculus enables inference by purely syntactic and mechanical transformations of formulas. A logical formula is “well formed” (WFF) if there are significant sequences based on a grammar of primary logic connectives $\{\neg, \vee, \wedge\}$. Brackets () are used to express connectives priority.

5 New Validation Process

In the following figure, we can see the scheme of the new validation process



The new process extends into several stages:

- Survey Project and Preparation (A – A1)
- Data Collection (B)
- Data Check (C – D – E)
- Data transfer to archives and to production units (G)

A Monitor control the data flow and records questionnaires status and overall completion status. Questionnaires check can fail only once, so unfilled and double failing units are separated and stored elsewhere. After that, validated questionnaires are ready for dissemination.

6 Conclusion

The project described above aims to improve the accuracy and consistency of the data through a new validation process, in a generalized perspective, which provides automatic procedures for checking the consistency of the data collected.

The system reduces monitors burden automating most tasks such as the interception of measurement errors, discontinuity of time series and other outliers.

Automatic interoperability with the municipalities contributes to increase efficiency of the overall process.

The result of the work is an integrated system of generalized services that will function on a formal level useful in many different contexts.

With regard to the production of official statistics, the system will make it possible to focus the work of researchers on the secondary level of data validation in relation to the trend and consistency aspects of the information collected.

References

1. Adamo, D., Costanzo, L., Buzzi, L., Laganà, A., Garozzo, S., Greco, V.: Principali fattori di pressione sull'ambiente nelle città italiane – Anno 2018, In: Adamo, D., Costanzo, L. (eds.) Istat, Territori, Letture statistiche, (ISBN: 978-88-458-2037-3) (2020)
2. Adamo, D., Bianchi, G., Mongelli, L.: Re-engineering environmental data collection in cities. RIEDS – Rivista Italiana di Economia Demografia e Statistica **78**(1), 2024 (2024)
3. Adamo, D.: Goal 11 – Make cities and human settlements inclusive, safe, resilient. In Adamo, D., (eds.: 2022 SDGs Report. Statistical information for 2030 Agenda in Italy, Istat, Territori, Letture statistiche (ISBN 978-88-458-2101-1) (2023)
4. Bianchi, G., Bruni, R.: Formal procedure for finding contradictions into a set of rules. Appl. Math. Sci. **6**(126), 6253–6271 (2012)
5. Bianchi, G., Filippini, R., Lipsi, R.M., Scalfati, F., Pezon, e A.: An overview of the editing and imputation process of the 2018 Italian Permanent census. In: Workshop on Statistical Data Editing, United Nations Economic Commission for Europe, Geneva, 31 Aug–4 Sep 2020
6. Bianchi, G., Manzari A., Reale, A.: An overview of editing and imputation methods for the next italian censuses (Key Invited Paper). In: Conference of European statisticians, Work Session on Statistical Data Editing, Geneva, 13–15 May 2008 (2008)
7. Bosch, P., Buchele, M., Gee, D.: Environmental Indicators: Typology and Overview. European Environment Agency, Copenhagen, Denmark (1999)
8. Chandru, V., Hooker, J.N.: Extended horn sets in propositional logic. J. ACM **38**, 205–221 (1991)
9. Torelli, R.: GINO++: A generalized system for web surveys. In: Proceedings of Statistics Canada Symposium 2011 (2011)



Detecting Circadian Gene Expressions via Bayesian Analysis: An Application to the *Arabidopsis Thaliana* Dataset

Silvia Montagna¹(✉), Amir Khorrami Chokami², and Surya T. Tokdar³

¹ Dipartimento di Economia “Marco Biagi”, Università degli Studi di Modena e Reggio Emilia, via Berengario 51, 41121 Modena, Italy

silvia.montagna@unimore.it

² Dipartimento di Matematica e Informatica, Università di Cagliari, Via dell’Ospedale 72, 09124 Cagliari, Italy

³ Department of Statistical Science, Duke University, Durham, NC 27708, USA

Abstract. In genomic applications, there is often interest in identifying genes whose time-course expression trajectories exhibit periodic oscillations with a period of approximately 24 h (circadian genes). While it is natural to expect that the expression of gene i at time j might depend to some degree on the expression of the other genes measured at the same time, widely-used rhythmicity detection techniques do not accommodate for the potential dependence across genes. We develop a Bayesian approach for periodicity identification that explicitly takes into account the complex dependence structure across time-course trajectories in gene expressions. The methodology is applied to a plant gene expression dataset.

Keywords: Bayesian Fourier analysis · high-dimensional data · circadian rhythms · gene expressions

1 Introduction

Circadian rhythms are cycles of biological activity based on a 24-hour period which allow organisms to anticipate and adapt to predictable daily oscillations in the environment [4]. Circadian rhythms are present in almost all plants and animals. In plants, circadian rhythms play a role in the regulation of plant metabolic pathways, such as photosynthesis and carbon metabolism, in the regulation of developmental processes and signalling pathways, such as defence responses. In humans, blood pressure, hormone production, metabolism and other biological cycles are clock-regulated. Disruptions to the circadian rhythms have been linked to a variety of pathologies, for example cancer, psychiatric disorders and neurodegenerative diseases in humans [4].

Circadian rhythms are controlled by the circadian clock, namely, a network of mutually interacting genes controlling the timing of many physiological processes. The interest is in identifying such genes through examination of their

expression levels or “transcripts”. Several authors have proposed methods for periodicity identification in biomedical research over the last couple of decades. [5] provide a comprehensive review of the main existing techniques commonly used for circadian rhythm detection, and evaluate their accuracy and reproducibility on various empirical datasets. As a separate line of research, several model-based clustering algorithms have been proposed in both the classical and Bayesian framework (e.g., [3]).

A key assumption in the approaches above is that of independence across genes. Although practical from a computational perspective, the independence assumption is often too strong to be realistic in many applications. In this paper, we propose a Bayesian approach that identifies periodic signals in gene expression profiles while accounting for dependence in the functional data. Specifically, the true underlying signal for each transcript is decomposed into a series expansion of sine and cosine (Fourier) waves. Conditional dependence across genes at each time point is accommodated via a latent factor framework. Dimensionality reduction and sparsity are induced through careful modelling of the latent factors as well as the Fourier basis coefficients.

The rest of the paper is organised as follows. Section 2 outlines the methodology. Section 3 discusses priors and posterior inference for detecting circadian genes. The proposed approach is tested on the *Arabidopsis thaliana* [3] dataset (Sect. 4). Conclusions are drawn in Sect. 5.

2 Methods

We consider data in the form of a $p \times T$ matrix $\mathbf{Y} = \{y_{ij}\}$, where entry y_{ij} denotes the expression level for gene i at time t_j , for $i = 1, \dots, p$, and with p denoting the total number of genes. In circadian microarray studies, data are typically collected over two complete circadian cycles and the sampling rate is usually either two ($T = 24$) or four hours ($T = 12$). We assume that the y_{ij} ’s are error-prone measurements of an underlying smooth true trajectory:

$$y_{ij} = f_i(t_j) + \nu_{ij}. \quad (1)$$

Suppose that the de-trended and centred true signal for gene i at time t_j , $f_i(t_j)$, can be represented as:

$$f_i(t_j) = \sum_{m=1}^q [\theta_{i,2m-1} b_{2m-1}(t_j) + \theta_{i,2m} b_{2m}(t_j)] = \sum_{m=1}^q \boldsymbol{\theta}_{i,m} \mathbf{b}_m(t_j) = \mathbf{b}_j^\top \boldsymbol{\theta}_i,$$

where $\boldsymbol{\theta}_{i,m} = (\theta_{i,2m-1}, \theta_{i,2m})$ and $\mathbf{b}_m(t_j) = [b_{2m-1}(t_j), b_{2m}(t_j)]^\top$, for $m = 1, \dots, q$. The vector $\mathbf{b}_j^\top = [b_1(t_j), b_2(t_j), \dots, b_{2q-1}(t_j), b_{2q}(t_j)]$ represents a set of $2q$ fixed basis functions evaluated at time t_j . The natural basis choice for the space of periodic functions is the Fourier basis:

$$\mathbf{b}^\top(t) = \left[\sin\left(\frac{2\pi}{\omega_1} t\right), \cos\left(\frac{2\pi}{\omega_1} t\right), \dots, \sin\left(\frac{2\pi}{\omega_q} t\right), \cos\left(\frac{2\pi}{\omega_q} t\right) \right],$$

where $\{\omega_m\}_{m=1}^q$ denotes the periodicity of the signal and t is time represented by a unit-interval increase. The q period lengths w_m are assumed known and fixed. Since there are 13 time points per transcript in the *Arabidopsis thaliana* dataset (Sect. 4), we can use up to six sine/cosine pairs of harmonics.

The term ν_{ij} in Equation (1) models the deviation between y_{ij} and the underlying smooth profile. To accommodate for the potential dependence across genes at time j , we adopt a factor model representation:

$$\boldsymbol{\nu}_j = \mathbf{\Lambda} \boldsymbol{\eta}_j + \boldsymbol{\epsilon}_j, \quad (2)$$

with $\boldsymbol{\nu}_j = [\nu_{1j}, \dots, \nu_{pj}]^\top$, $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_p]^\top$ is a $p \times k$ factor loading matrix with elements $\{\lambda_{ih}\}_{i=1, \dots, p; h=1, \dots, k}$, $\boldsymbol{\eta}_j = (\eta_{1j}, \dots, \eta_{kj})^\top$ is $k \times 1$ vector of latent factors at time j , and $\boldsymbol{\epsilon}_j$ is a residual error. The full model for gene i at time t_j is:

$$y_{ij} = \mathbf{b}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\lambda}_i \boldsymbol{\eta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_i^2), \quad (3)$$

where the first term $\mathbf{b}_j^\top \boldsymbol{\theta}_i$ captures periodic oscillations whereas the second term $\boldsymbol{\lambda}_i \boldsymbol{\eta}_j$ captures across-genes dependence (if present).

Hereafter we follow standard practice and assign a Normal prior to the latent factors at time t_j , $\boldsymbol{\eta}_j \sim N_k(\mathbf{0}, \mathbf{I})$. Genes are assumed to be independent given the latent factors, and dependence among genes is induced by marginalising over the distribution of the factors. Therefore, marginally $\mathbf{y}^{(j)} \sim N(\boldsymbol{\Theta} \mathbf{b}_j, \mathbf{\Lambda} \mathbf{\Lambda}^\top + \boldsymbol{\Sigma})$, with $\mathbf{y}^{(j)} = (y_{1j}, \dots, y_{pj})^\top$, $\boldsymbol{\Theta}$ is the $p \times 2q$ matrix of basis function coefficients, and $\boldsymbol{\Sigma} = \text{Diag}(\sigma_1^2, \dots, \sigma_p^2)$. In practical applications involving moderate to large p , the number of factors k is typically much smaller than p , thus inducing a sparse characterisation of the unknown covariance matrix $\mathbf{\Lambda} \mathbf{\Lambda}^\top + \boldsymbol{\Sigma}$.

3 Prior Elicitation and Identification of Circadian Genes

Inference on parameters $\boldsymbol{\theta}_i$ is the primary interest of our work. Recall that $\boldsymbol{\theta}_{i,m} = (\theta_{i,2m-1}, \theta_{i,2m})$, where $\theta_{i,2m-1}$ is the coefficient of the $(2m-1)$ -th sine basis and $\theta_{i,2m}$ is the coefficient of the $2m$ -th cosine basis, both harmonics of period w_m . To allow for the correct identification of gene i 's periodicity, we need to switch on/off $(\theta_{i,2m-1}, \theta_{i,2m})$ jointly. We assume a latent threshold prior (LTP) [6]:

$$\boldsymbol{\theta}_{i,m} = \tilde{\boldsymbol{\theta}}_{i,m} \mathbb{1}(\|\tilde{\boldsymbol{\theta}}_{i,m}\| \geq \varpi_{i,m}), \quad (4)$$

where $\varpi_{i,m}$ is a latent threshold. The idea behind (4) is that the m -th pair of sine/cosine coefficients is shrunk to zero when their (Euclidian) norm falls below a m -th- (and gene-) specific threshold. Further, we model $\tilde{\boldsymbol{\theta}}_i = \{\tilde{\boldsymbol{\theta}}_{i,m}\}_{m=1}^q$ as:

$$\tilde{\boldsymbol{\theta}}_i = \mathbf{W} \boldsymbol{\lambda}_i^\top + \boldsymbol{\delta}_i \quad \text{and} \quad \boldsymbol{\delta}_i \sim N_{2q}(\mathbf{0}, \mathbf{I}), \quad (5)$$

where $\boldsymbol{\lambda}_i$ is the vector of factor loadings for gene i as in Eq. (3), and $\mathbf{W}$ is a $2q \times k$ matrix such that $\mathbf{W}_j^\top \sim N_k(\mathbf{0}, \mathbf{I})$, $j = 1, \dots, 2q$. Elicitation is completed by placing conditionally conjugate priors on all remaining model parameters, e.g., Gamma priors on precisions, the multiplicative Gamma process shrinkage

prior [1] on the loadings and a uniform prior on the latent thresholds. Posterior inference proceeds via MCMC with conditionally conjugate updating steps.

The LTP eases the identification of circadian genes. Indeed, we estimate the posterior probability of a gene being circadian by counting the proportion of posterior samples for which the coefficients of the 24-hours sine/cosine pair (e.g., $(\theta_{i,2q-1}, \theta_{i,2q})$) are *not* shrunk to zero whilst all the remaining coefficients are switched off:

$$P(\text{Gene } i \text{ is circadian}) = \frac{1}{TS} \sum_{g=1}^{TS} \mathbb{1} \left(\left\{ \theta_{i,l}^{(g)} \right\}_{l=1}^{2q-2} \equiv \mathbf{0} \text{ and } \left\{ \theta_{i,2q-1}^{(g)}, \theta_{i,2q}^{(g)} \right\} \neq \mathbf{0} \right), \quad (6)$$

where g denotes the iteration number and TS denotes the total number of thinned posterior samples post burn-in. Similarly, the framework above can be used for inference on phase and amplitude of the signals, or for detecting whether a gene exhibits periodicity other than 24-hours.

4 Analysis

We apply our method to the *Arabidopsis Thaliana* dataset [3]. Eight-day-old Columbia seedlings grown under 12-hours-light/12-hours-dark cycles were transferred to constant light at 22°. Plant samples were harvested at 13 time points covering two circadian cycles in 4 h intervals, starting 26 h after the last dark-light transition. Here $p = 22810$ genes and $T = 13$ time points.

[2] reports 26 known clock-associated genes in *Arabidopsis*. Among these genes are CCA1 (Circadian Clock Associated 1) and LHY (Late Elongated Hypocotyl), which function synergistically in regulating circadian rhythms of *Arabidopsis*, TOC1 (Timing of Cab Expression 1), which contributes to the plant fitness (carbon fixation, biomass), and ELF4 (Early Flowering 4), which accounts for sustained rhythms in the absence of daily light/dark cycles. We use the 26 well-known circadian genes as a benchmark to evaluate our approach. We compare the proposed approach with its independent version (that is, omitting the latent factor component from Eq. (3)) and JTK_Cycle [4] (chosen for comparison based on performance evaluations presented in [5]). The rankings of the 26 known clock genes are reported in Table 1. All the algorithms were able to identify most of the known clock genes from among their top 25% ranked candidates, and we observe improved performance in terms of placing more genes in the top 1% and 25%. The independent version of our model also performs well. This is likely due to the borrowing of information in modelling the trajectories induced by the Bayesian framework. Estimated trajectories for four clock genes are represented in Fig. 1.

We finally remark that we have also evaluated our model on simulation studies, where we could compare its performance to that of competitors in settings of both dependence and independence across synthetic trajectories. These studies show that our construction gives improved performance in identifying rhythmic curves over widely-used rhythmicity detection techniques in both settings. Results are omitted here due to space limitations.

Table 1. Summary of rankings of 26 known clock genes in the *Arabidopsis Thaliana* genome. Genes were ranked by estimated posterior circadian probability for the proposed approach (below Dep.LF) and its independent version (IndepV); by p -value for JTK cycle.

Method	Top 1%	Top 5%	Top 10%	Top 25%	Top 60%
Dep. LF	4	8	14	23	26
IndepV	2	6	13	22	25
JTK cycle	1	8	12	17	22

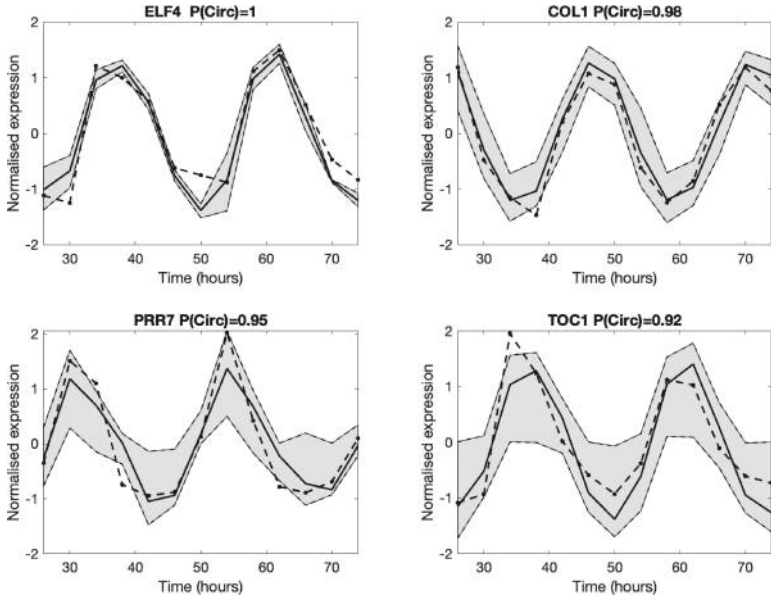


Fig. 1. Four known clock genes in the *Arabidopsis* dataset ranking top by estimated posterior probability of being circadian. The black dashed trajectory connects the true expression levels (dots), the solid black line represents the estimated posterior mean trajectory and the shaded grey area represents point-wise 95% credible intervals around the estimated posterior mean trajectory.

5 Conclusions

In this manuscript, we presented a Bayesian method for periodicity detection. The method employs a Fourier basis expansion coupled with a variable selection prior on the basis coefficients to model gene expression trajectories and identify circadian genes. The core statistical contribution consists in accommodating for the potential dependence in the trajectories via latent factor modelling. We apply our technique to a well studied gene expression dataset, and its performance is line with (if not better than that) of a widely-used rhythmicity detection technique that does not directly accommodate for dependence across trajectories.

The results presented in this manuscript should be considered as preliminary results, and various extensions could be considered. For example, in animal (e.g., mice) studies, mice could be given a stimulus at the beginning of the experiment. The stimulus may produce local deviations in expression levels, and these deviations may manifest at different times across genes and last for a different amount of time. Deviations could be accommodated for by including an additional local bases decomposition term in Eq. (3). Further, other priors for the basis coefficients or for the latent factor model could also be considered. One drawback of accommodating dependence across curves is the increased computational demand, although the latent factor representation is a convenient way of doing so. Gene-specific updating steps can also be parallelised to speed up posterior computation. Although developed for detecting circadian genes, we finally remark that the approach can be applied to any dataset where the inferential goal is that of detecting periodicity of curves. Indeed, we are currently testing our model to a novel EMG dataset for the identification of periodic bursts in spinae muscles of violin players. Extensions and results will be collected in future research.

References

1. Bhattacharya, A., Dunson, D.B.: Sparse Bayesian infinite factor models. *Biometrika* **98**, 291–306 (2011)
2. Dodd, A.N., et al.: The Arabidopsis circadian clock incorporates a cADPR-based feedback loop. *Science* **318**, 1789–1792 (2007)
3. Edwards, K.D., et al.: Flowering locus C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. *Plant Cell* **18**, 639–650 (2006)
4. Hughes, M.E., Hogenesch, J.B., Kornacker, K.: JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J. Bio. Rhythms* **25**, 372–380 (2010)
5. Mei, W., Jiang, Z., Chen, Y., Chen, L., Sancar, A., Jiang, Y.: Genome-wide circadian rhythm detection methods: systematic evaluations and practical guidelines. *Brief. Bioinf.* **22**, 1–13 (2021)
6. Nakajima, J., West, M.: Bayesian analysis of latent threshold dynamic models. *J. Busin. Econ. Stat.* **31**, 151–164 (2013)
7. Straume, M.: DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Meth. Enzym.* **383**, 149–166 (2004)



The Benefit of Social Capital Resources for Contrasting Ethnic Bullying at School: A Multilevel Study

Maria Francesca Morabito^(✉) , Serena Verbena , Valentina Tocchioni ,
and Benedetta E. Palladino

University of Florence, Florence, Italy
mariafrancesca.morabito@unifi.it

Abstract. This study explores the multifaceted dynamics of ethnic victimization at school. Previous research has focused on the harmful consequences among victims of bullying; through this study, we seek to shed light on the understudied role of various social capital factors in mitigating the harm of ethnic bullying in terms of victims' emotive symptoms. We employ a large set of data on Italian high school students collected in 2023. Through a multilevel analysis, we find that individual social capital factors are associated with emotive symptoms among ethnically bullied students. Most social capital factors reduce the severity of emotional symptoms among victims, with peer connectedness and teachers' intervention in case of bullying identified as the most important in this respect. Instead, rule clarity and those factors that promote the expression of victims' emotions and enhance their consciousness amplify victims' emotional symptoms. Interestingly, at the school level, only a few of them emerge as significant. These findings can lay the groundwork for educational policies and interventions aimed at promoting students' adaptation and well-being through the maximization of social capital.

Keywords: ethnic bullying · high school · multilevel models · Italy

1 Introduction

This study aims to explore the multifaceted dynamics of ethnic victimization at school by delving into the role of a range of social capital-related resources. Ethnic bullying represents a widespread and enduring challenge, particularly within educational environments [11]. Previous research has extensively focused on the harmful consequences among victims, while comparatively little attention has been devoted to evaluating the role of social capital-related resources for students who suffer from being bullied based on their culture or ethnicity. The absence of teachers' support in addressing (offline) bullying episodes has been shown to exacerbate cyberbullying [8], as well as the protective function of the school climate on victims of bullying has been documented at both the individual [12] and the school level [2]. In this work, we intend to examine the extent to which individual- and school-level social capital factors are associated with emotive

symptoms among students who suffered from being bullied because of an ethnic background in Italy. In this respect, we want to ascertain if social capital resources may play a protective role for students of diverse ethno-cultural origins who suffer from being bullied at school for their race or ethnicity.

2 Data and Method

We employ a large dataset gathered from a high school students survey dedicated to fighting bullying and cyberbullying and funded by the Italian Ministry of Education. The survey was conducted at the end of the school year (i.e., May-June) in 2023 ($N = 185,063$) on high school students from 9th to 13th grade. We select those participants aged between 14 and 20 years, who reported being bullied (occasionally or systematically) based on their ethnicity and culture ($N = 17,202$, corresponding to 9.3% of the total sample, dislocated in 585 high schools).¹ To examine the role of social capital-related factors at the individual and the school level in mitigating the harm of being ethnically bullied in terms of emotional symptoms, we apply a multilevel regression model with students nested within schools. By including school-level information, we aim to test whether the school where a student is enrolled plays a role in such an association.

The response variable is the emotional symptoms scale proposed by Goodman [5] in the Strengths and Difficulties Questionnaire (SDQ). The validity and reliability of the SDQ scales are reported by scholars [6, 7, 10]. The scale capturing emotional symptoms of (ethnically bullied) students is constructed by aggregating the responses to survey items related to the emotivity sphere and ranges from 0 to 1. These items are: “I get a lot of headaches, stomachaches or sickness”, “I worry a lot”, “I am often unhappy, downhearted or tearful”, “I am nervous in new situations. I easily lose confidence”, and “I have many fears. I am easily scared”.

As for the key explanatory variables, regarding social capital resources at the individual level, we consider: (a) the four WHITS scales introduced and validated by Aldridge and Ala'I in [1] to measure the student's perception of the school climate; (b) five scales specifically referring to teachers' responses to bullying (TRBQ) proposed by Campaert and colleagues and revised by Nappa and colleagues [4, 8]; and (c) two indices of the school context. Specifically, the four scales about the perceived school climate (a) refer to: (1) *reporting and seeking help*, (2) *rule clarity*, (3) *peer connectedness*, and (4) *teacher support*. Regarding the five TRBQ scales (b), we use measures of teachers' (5) *intervention* in addressing bullying episodes, (6) *disciplinary methods* they adopt in these instances, (7) *victim support*,² (8) *group discussion* about bullying, and the ability of (9) *mediation* by teachers in case of conflicts among students. In addition, as measures of the school context (c), we consider the student's (10) *knowledge of consequences* of bullying episodes, and the student's perception of (11) *school safety*. Among others, Riekie

¹ Data from schools where less than 10 students reported being ethnically bullied are excluded from the analysis (4.3% of the selected sample).

² *Teacher support* (4) and *victim support* (7) differ from each other. The former refers to the student's perceptions of the overall support provided by teachers to all students, whereas the latter specifically pertains to the student's perceptions of the adequacy of support provided by teachers to victims of bullying.

and colleagues [9] used the scales (1)–(4) to assess the (positive) association between school climate and the mental health of students; Burger and colleagues [3] employed the scales (5)–(9) to demonstrate that teachers' responses to bullying intervene in the construction of bullying-related roles.

We include the means of the variables on (12) *reporting and seeking help* at school, and (13) *teacher support* in each school as school-level measures. We tested the inclusion of other social capital variables at the school level as means of individual-level social capital resources (such as (3) *peer connectedness*, and teachers' (5) *intervention* in case of bullying), but they were not statistically significant in association with emotional symptoms of ethnic-based victims, and we dropped them from the final model.

All individual- and school-level scales are normalized to vary within the same range (0–1) and included as continuous variables. Finally, we control for the sex and age of the student, the school class attended, the educational track followed, the region where the school is located, and the percentage of female students interviewed within the school, which are all covariates reported in the literature as being associated with the harm of ethnic bullying.

3 Results

Results from the multilevel regression are reported in Table 1. Regarding school climate (a), students who are bullied because of their ethnicity and are well acquainted with the school rules tend to report more negative emotional symptoms. Instead, when ethnic-based victims perceive that students are adequately supported by teachers and have a better sense of connection with their peers, their emotional symptoms are alleviated. Teachers' responses to bullying, namely, the TRBQ scales (b), have an important role in shaping the emotional symptoms of ethnically bullied students: a higher capability of intervention and mediation by teachers in instances of bullying is associated with fewer emotional symptoms, while negative emotional symptoms increase when bullied students feel that teachers are more capable of supporting bullying victims and when bullying episodes are adequately discussed at school. This may be due to an increased inclination of students to express discomfort in environments where speaking openly about bullying is destigmatized and where they are confident of receiving support. Considering the school context (c), being aware of the consequences of bullying for bullies and feeling safe at school function as protective factors for emotional symptoms among students who suffered from being bullied based on their ethnicity.

At the school level, being helped in reporting cases of bullying increases the likelihood of reporting more emotional symptoms by ethnically bullied students, whereas adequate teachers' support is associated with fewer emotional symptoms. Among control variables, female students and students who do not recognize themselves in any of the two sexes feel more negative emotional symptoms than male students. Moreover, students attending lyceum report more emotional symptoms than those attending professional or technical institutes. Finally, students bullied because of their ethnicity show more emotional symptoms as they progress through the school grades, which could suggest that ethnic bullying episodes, if not adequately contrasted but repeated over the years, worsen victims' emotional status over time.

Table 1. Multilevel regression model on emotional symptoms among bullied students.

Variable	Coefficient	Standard Error	P-value
INDIVIDUAL LEVEL			
Social capital factors			
(1) Reporting and seeking help	0.007	0.011	0.541
(2) Rule clarity	0.110	0.010	0.000
(3) Peer connectedness	−0.118	0.008	0.000
(4) Teacher support	−0.023	0.010	0.027
(5) TRBQ Intervention	−0.150	0.009	0.000
(6) TRBQ Disciplinary methods	0.016	0.011	0.136
(7) TRBQ Victim support	0.035	0.013	0.008
(8) TRBQ Group discussion	0.027	0.010	0.010
(9) TRBQ Mediation	−0.051	0.012	0.000
(10) Knowledge of consequences	−0.020	0.007	0.004
(11) School safety	−0.053	0.007	0.000
<i>Other individual-level covariates</i>			
Educational track (ref: Lyceum)			
Technical institute	−0.014	0.005	0.008
Professional institute	−0.016	0.006	0.010
Sex (ref: Male)			
Female	0.204	0.004	0.000
Other	0.122	0.008	0.000
School grade (ref: 9)			
10	0.015	0.006	0.011
11	0.024	0.007	0.002
12	0.034	0.009	0.000
13	0.030	0.010	0.004
Age	−0.077	0.032	0.015
Age ²	0.002	0.001	0.023
SCHOOL LEVEL			
Social capital factors			
(12) Reporting and seeking help	0.148	0.051	0.004
(13) Teacher support	−0.168	0.046	0.000
<i>Other school-level covariates</i>			

(continued)

Table 1. (continued)

Variable	Coefficient	Standard Error	P-value
Region			
North-East	−0.009	0.006	0.158
Center	−0.004	0.007	0.539
South	−0.003	0.006	0.572
Islands	−0.017	0.009	0.051
Percentage of female students	0.040	0.013	0.002
CONSTANT	1.130	0.258	0.000
Individuals	17,202		
Schools	585		

Note: The variable “Educational track” is included as a first-level variable since within the same school institution (the second level), different educational tracks may exist.

4 Conclusions and Further Developments

In our analyses, it emerges clearly how individual social capital factors are associated with emotive symptoms among students who suffered from being bullied because of their ethnicity. In particular, rule clarity, peer connectedness, and teachers’ intervention in case of bullying play a major role in shaping the emotional symptoms of ethnically bullied students. These associations are mostly negative except when considering aspects that may encourage the expression of victims’ emotions (such as victims’ support, ease in finding help, and openness to dialogue on bullying) and increase their consciousness (like rule clarity).

The contribution of this study is twofold. On the one hand, we add to the previous research on ethnic bullying, in particular evaluating the role of social capital-related resources for students who suffer from being bullied based on their culture or ethnicity. In this respect, we intend to implement other modeling strategies, too, such as multilevel latent class analysis, to deepen the understanding of this relationship. More specifically, we aim to identify latent classes of students, thus obtaining specific profiles based on different combinations of behaviors of ethnically bullied students; and explore the effects of individual-level, school-level, and contextual-level social capital resources on these behaviors within each latent class. Even more, we intend to perform heterogeneity analyses by being bullied occasionally or systematically to check if the effects of social capital resources on emotional symptoms vary between the two groups.

On the other hand, it may offer practical guidance for educators and policymakers seeking to promote students’ adaptation and well-being. For example, it seems to highlight how some factors seem fundamental to contrast the harmful consequences of ethnic bullying, such as teachers’ interventions. Even more, it seems to suggest that precocious action in the first grades of high school could be essential to contrast the worsening of negative emotional symptoms during the school years. Indeed, the findings of this study

can lay the groundwork for educational policies and interventions aimed at enhancing learning environments through the maximization of social capital.

References

1. Aldridge, J., Ala'I, K.: Assessing students' views of school climate: developing and validating the what's happening in this school? (WHITS) questionnaire. *Improv. Sch.* **16**(1), 47–66 (2013). <https://doi.org/10.1177/1365480212473680>
2. Aldridge, J.M., McChesney, K., Afari, E.: Relationships between school climate, bullying and delinquent behaviours. *Learn. Environ. Res.* **21**, 153–172 (2018). <https://doi.org/10.1007/s10984-017-9249-6>
3. Burger, C., Strohmeier, D., Kollerová, L.: Teachers can make a difference in bullying: effects of teacher interventions on students' adoption of bully, victim, bully-victim or defender roles across time. *J. Youth Adolesc.* **51**, 2312–2327 (2022). <https://doi.org/10.1007/s10964-022-01674-6>
4. Campaert, K., Nocentini, A., Menesini, E.: The efficacy of teachers' responses to incidents of bullying and victimization: the mediational role of moral disengagement for bullying. *Aggress. Behav.* **43**(5), 483–492 (2017). <https://doi.org/10.1002/ab.21706>
5. Goodman, R.: The strengths and difficulties questionnaire: a research note. *J. Child Psychol. Psychiatry* **38**(5), 581–586 (1997). <https://doi.org/10.1111/j.1469-7610.1997.tb01545.x>
6. Goodman, R., Meltzer, H., Bailey, V.: The strengths and difficulties questionnaire: a pilot study on the validity of the self-report version. *Eur. Child Adolesc. Psychiatry* **7**(3), 125–130 (2003). <https://doi.org/10.1007/s007870050057>
7. Muris, P., Meesters, C., van den Berg, F.: The strengths and difficulties questionnaire (SDQ). *Eur. Child Adolesc. Psychiatry* **12**(1), 1–8 (2003). <https://doi.org/10.1007/s00787-003-0298-2>
8. Nappa, M.R., Palladino, B.E., Nocentini, A., Menesini, E.: Do the face-to-face actions of adults have an online impact? The effects of parent and teacher responses on cyberbullying among students. *Eur. J. Dev. Psychol.* **18**(6), 798–813 (2021). <https://doi.org/10.1080/17405629.2020.1860746>
9. Riekke, H., Aldridge, J.M., Afari, E.: The role of the school climate in high school students' mental health and identity formation: a South Australian study. *Br. Educ. Res. J.* **43**(1), 95–123 (2017). <https://doi.org/10.1002/berj.3254>
10. Van Roy, B., Veenstra, M., Clench-Aas, J.: Construct validity of the five-factor strengths and difficulties questionnaire (SDQ) in pre-, early, and late adolescence. *J. Child Psychol. Psychiatr.* **49**(12), 1304–1312 (2008). <https://doi.org/10.1111/j.1469-7610.2008.01942.x>
11. Wolke, D., Copeland, W.E., Angold, A., Costello, E.J.: Impact of bullying in childhood on adult health, wealth, crime, and social outcomes. *Psychol. Sci.* **24**(10), 1958–1970 (2013). <https://doi.org/10.1177/0956797613481608>
12. Yoneyama, S., Rigby, K.: Bully/victim students and classroom climate. *Youth Stud. Aust.* **25**, 34–41 (2006)



Computing Multidimensional Composite Indicators for Small Areas in Presence of Missing Variables: A Data Integration Approach

Angelo Moretti¹✉, Alejandra Arias-Salazar², and Natalia Rojas-Perilla³

¹ Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands
a.moretti@uu.nl

² University of Costa Rica, San Jose, Costa Rica

³ United Arab Emirates University, Abu Dhabi, UAE

Abstract. In this article, we assess data integration methods for estimating composite indicators for small areas, where some single indicators are completely missing. Focusing on a multidimensional poverty index, with certain variables missing from the population Census, we propose two approaches using an auxiliary sample survey. One approach employs a generalized linear mixed model, while the other employs a two-step imputation technique. We evaluate these approaches through simulation studies, and an application based on the Colombia's 2018 Great Integrated Household Survey as a case study.

Keywords: composite indicators · GLMM · statistical matching · mass imputation

1 Introduction to the Problem and Notation

Composite indicators offer a nuanced understanding of complex phenomena like poverty and well-being. Unlike single-dimensional measures e.g., income, multidimensional indices consider various dimensions like education, health, and housing simultaneously, providing a more holistic view of deprivation and informing effective policymaking [1, 2]. However, computing composite indicators often requires integrating data from multiple sources, presenting challenges when some variables are missing in the primary data source. We propose a data integration framework for small area estimation (SAE) context, focusing on missing dimensions represented by binary variables. A two-step imputation-based technique [3] and a model-based approach using a generalized linear mixed model [4] are considered. Illustrated using Colombia as a case study, these methods demonstrate effectiveness in addressing missing indicators issues and enhancing estimation accuracy for composite indicators. The proposed techniques can be applied in similar data structures and scenarios, offering valuable insights for policymakers and researchers in the field. We focus on a Multidimensional Poverty Index (MPI) following the characteristics implemented in [4]. It requires $t = 1, \dots, T$, with $T = 8$ indicators

which are measured as deprivations: $y_{dit} = 1$ if the individual i in are d is deprived in the indicator t and 0 otherwise. Notice that its construction requires information for each individual $i = 1, \dots, N_d$ in $d = 1, \dots, D$ domains, where N_d denotes the population size of d . We assume that the D domain are non-overlapping. The MPI for the domain d is computed as the product of two measures: “H”, the headcount ratio or incidence of multidimensional poverty and “A”, the intensity [5]:

$$MPI_d = H_d \cdot A_d \quad (1)$$

With $H_d = N_d^{-1} \sum_{i=1}^{N_d} I(q_{di} > z)$, where $I(\cdot)$ will be 1 if q_{di} is larger than a threshold z . Similar to [4], we define $z = 0.4$ (we refer to [4] for details due to space constraints). q_{di} is a weighted sum $q_{di} = \sum_{t=1}^T \Delta_{y_{dit}}$ where the first six indicators $t = 1, \dots, 6$ receive a weight of $\Delta = 0.1$ and $t = 7$ and $t = 8$ have a weight of $\Delta = 0.2$. Intensity “A”, is computed as the average deprivation score, considering only the individuals classified as multidimensional poor, i.e., $A_d = \Phi_d^{-1} \sum_{i=1}^{\Phi_d} q_{di}$, with $\Phi_d = \sum_{i=1}^{N_d} I(q_{di} > z)$. As a case-study for the application in Sect. 4 and evaluations in Sect. 3, we use Colombian data because it provides an important real scenario: on the one hand, there is a household survey (Large Integrated Household Survey) containing all the required information to compute the MPI, however, MPI estimates are not available for smaller areas or population groups due to very small sample sizes. On the other hand, the Census would be the perfect tool to provide quality MPI values for specific groups, but it has two missing indicators which makes its computation impossible. For this specific case study we set 6 existing indicators in the Census: poor housing materials, overcrowding, lack of drinking water, lack of sanitation, lack of internet service, lack of electricity. The missing indicators in the Census are: unfinished education and unemployment or insufficient employment-related income or pension. Let us consider a Census file, denoted by C , related to a target population U , with dimension N . Here, two binary variables, with indexes $k = 1, \dots, K$, i.e., y_{dik} thus $K = 2$, are not collected, hence are missing for all the units $i \in C$. K denotes the non-collected variables. The total number of indicators is denoted by T , with $K < T$ and with indexes $t = 1, \dots, T$. Notice that we consider 2 missing indicators only, however our approaches can be extended to a general case. In addition, we assume the Census records information on a wide range of P socio-economic characteristics, as well as deprivation variables, which we denote by $\mathbf{x}_{di} = (x_{di1}, \dots, x_{diP})$. Let us now assume, as in our case study, that there is a random sample, denoted by s , with dimension $n < N$, but where variables y_{di1} and y_{di2} $i \in s$ are collected. In addition, we assume that there are some variables which are common to the Census and sample. Here, similarly to the Census, each unit i is nested into a small area d . This survey has an important drawback, i.e., it is highly unrepresentative for small areas with a large number of out-of-sample areas (unplanned domain issue). We intend to tackle the specific problem of producing MPI estimates for small areas where some important variables are missing in the Census data. For this, we frame the issue as a data integration problem, thus, we compare two integration methods.

2 Two Data Integration Approaches for Binary Variables

Our aim is to generate y_{dik} in $i \in C$ for the missing variables $k = 1, \dots, K$. The first strategy we propose is a two-step imputation approach (Imp) and it is in line to the one used in [3] in the context of augmenting a sample survey for the creation of a reference probability sample for non-probability sample weighting. First, we match each record of the sample to the Census using a constrained statistical matching based on the hot-deck [6]; thus, n records are imputed in y_{dik} for with $i \in C$ for K variables. We set the small area as donation classes to preserve spatial information. From this step, $N - n$ non-matches are produced; hence these are mass-imputed via Fractional Hot-Deck Imputation [3, 7]. The process will return a complete Census file C which will be used to compute the MPI. In the second strategy (GLMM), we assume a Bernoulli logit mixed model:

$$\text{logit}(\pi_{dik}) = \log\left(\frac{\pi_{dik}}{1 - \pi_{dik}}\right) = \mathbf{x}_{di}^T \beta_k + u_{dk} \quad (2)$$

This approach consists in the following steps: 1. Estimate (2) for the indicators of interest and obtain the model estimates for each variable k . 2. Run the following Monte Carlo procedure, with replications $l = 1, \dots, L$, for each k : i. using the Census predict the probability that each unit obtains the value 1 as follows: $\hat{\pi}_{dik}^{in(l)} = \frac{\exp(\mathbf{x}_{di}^T \hat{\beta}_k + \hat{u}_{dk})}{1 + \exp(\mathbf{x}_{di}^T \hat{\beta}_k + \hat{u}_{dk})}$, $i \in C$, ii. Obtain the Monte Carlo estimates $\tilde{y}_{dik}^{(l)}$ with $\tilde{y}_{dik}^{(l)} \sim \text{Bernoulli}(\hat{\pi}_{dik}^{in(l)})$, iii. With the available indicators in the Census and the $K = 2$ new generated indicators $\tilde{y}_{dik}^{(l)}$ computed $H_d^{(l)}$ and $A_d^{(l)}$. 3. Compute the Monte Carlo approximation of the MPI as follows: $\widehat{MPI}_d = L^{-1} \sum_{l=1}^L H_d^{(l)} \times L^{-1} \sum_{l=1}^L A_d^{(l)}$.

3 Model-Based and Design-Based Simulation Studies

To create the universe in the model-based simulation, we first select a random sample from the Economic Commission for Latin America and the Caribbean (ECLAC) version of the Large Integrated Household Survey of Colombia: we sample 10% of the households via simple random sampling without replacement selection within each municipality, resulting in the following dimensions: $N = 21837$, and $D = 220$. Since we want to study the sensitivity to model assumptions, we generate the indicators that are missing $y_{di1}^{(l)}$ and $y_{di2}^{(l)}$ with $l = 1, \dots, 500$ following model (2), but we correlate the random effect with $\rho = 0.5$, given that in real that these are strongly related. These will be considered as the ‘truth’. The independent variables in both models are: age (7 bands), sex, and ethnicity (3 categories), and we use the following model¹: $\beta_1 = (1.29, -0.39, -0.58, -0.41, -0.13, -1.88, -1.36, 0.12, 0.5)^T$, $\beta_2 = (0.9, -0.2, -0.22, -0.01, 0.15, -0.7, -2.05, 0.23, 0.51, 0.11)^T$, $\sigma_{u_1}^2 = 0.25$ and $\sigma_{u_2}^2 = 0.61$. We also consider two scenarios of model misspecification: outliers, by

¹ These are obtained by estimating two generalized linear mixed regression models on the survey data.

contaminating the random effects as product of two distributions as in [8] as follows $(1 - 0.3)\mathbf{u}_d + 0.3\mathbf{u}_d^*, \mathbf{u}_d^* \sim N\left(0, \begin{bmatrix} 6.25 & 4.75 \\ 4.75 & 15.25 \end{bmatrix}\right)$ and heteroskedasticity i.e., $\mathbf{u}_d^{\text{het}} = (1 - 0.35)\mathbf{u}_d + .65\bar{x}_d^{\text{ethnic}1}$, where $\bar{x}_d^{\text{ethnic}1}$ is the proportion of people in area d that do not belong to either an indigenous or African ethnic group. Notice that ethnicity is a variable used to generate the responses in the population. Then, we select a SRS without replacement within each area with $n_d = 10$ from each l^{th} population. Since we assume that the sample may be drawn at a different time period than the Census, or there might errors in the data editing, we perturb 35% of the sample records of the variables used in the data integration strategies which are: sex, age, housing ownership (2 categories), as well as the following binary material deprivation variables owning a TV, Internet, washing machine, fridge, heating, and computer. The perturbation involves taking values from the lag of the age and a reversed variable for the binary variables. In each replication, we integrate the sample to the Census considering using the strategies in Sect. 2. The following variables are used – for the FHDI: sex, age, department, urbanity, ethnicity, house tenancy, owning TV, Internet, washing machine, fridge, heating, computer; for the GLMM-education: department, sex, age, urbanity, ethnicity, homeless indicator, owning a holiday house, owning a computer, having air conditioning, owning a car and electricity; GLMM-employment: sex, age, department, indicator whether the person was born in the municipality, urbanity, homeless indicator, house tenancy, air conditioning, and electricity. The variable selection for the GLMMs was carried out via the LASSO. We calculate the Absolute Relative Bias (ARB) and Relative Root Mean Squared Error (RRMSE) on the MPI indicators. Table 1 presents the ARM and RRMSE of the MPI estimates computed with the two strategies and under non contaminated and contaminated scenarios². Imp performs better in terms of ARB for the non contaminated model; however, for the contaminated model some ARB values in Imp exceed the ones obtained via the GLMM. Nonetheless, we point out that the ARB values are still very small for both approaches. The same can be argued for the RRMSE. Figure 1 shows the ARB and RRMSE across the small areas in case of heteroskedasticity. In this case, it can be seen that Imp provides better quality estimates compared to the use of the GLMM across all the areas Table 1.

Table 1. Summary of ARB and RRMSE of GLMM and Imp estimators with non contaminated and contaminated (with outliers) random effects (model-based simulation).

	Not contaminated				Contaminated			
	ARB		RRMSE		ARB		RRMSE	
	GLMM	Imp	GLMM	Imp	GLMM	Imp	GLMM	Imp
Min.	0.000	0.000	0.003	0.000	0.000	0.000	0.003	0.000
1st Qu.	0.006	0.004	0.009	0.005	0.006	0.004	0.010	0.005
Median	0.012	0.009	0.013	0.008	0.012	0.007	0.012	0.009
Mean	0.015	0.012	0.014	0.009	0.014	0.012	0.013	0.013
3rd Qu.	0.020	0.016	0.018	0.012	0.021	0.014	0.016	0.018
Max.	0.066	0.059	0.035	0.025	0.063	0.087	0.031	0.043

² Due to space constraints, we only selected a subset of outputs that we produced in the simulation studies. Thus, we are showing here the most interesting and impactful.

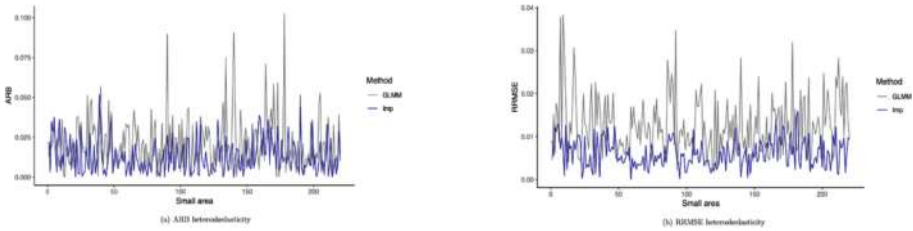


Fig. 1. ARB (a) and RRMSE of Imp and GLMM estimators (b) when heteroskedasticity is introduced (model-based simulation).

We also conducted a design-based simulation study; these are important since we can evaluate the performances of the methods in case of repeated samples drawn from a fixed real population, which does not follow the assumed model. We select a SRS without replacement for each area d with $n = 7$, and repeat this 500 times. Similarly to the model-based simulation study, we apply both approaches on each sample, repeat the simulations 500 times, and evaluate the same quality measures on the MPI. Figure 2 shows the ARB and RRMSE of the MPI estimates across areas. It can be seen that Imp provides better quality estimates across all the areas, and GLMM shows some peaks.

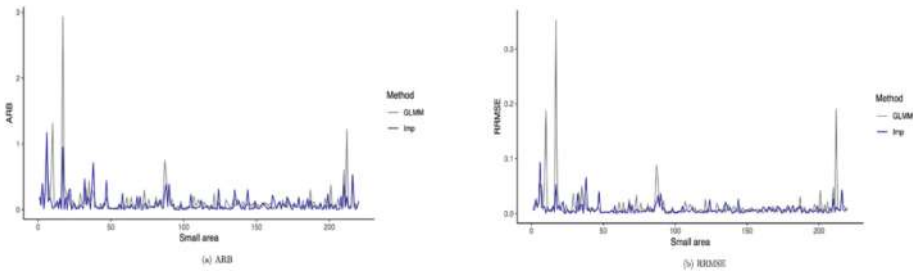


Fig. 2. ARB (a) and RRMSE of Imp and GLMM estimators for the design-based simulation.

4 Example Based on Real Data

We assume that the survey data (household file) is the Census file, and we draw a random sample from it with simple random sampling, setting municipalities (second administrative units) as strata. Defining a closer sampling design to the one used by the DANE is not possible due to a lack of information such as primary sampling units. The dimensions are the following ones: $N = 230827$, $D = 438$, with a sample size equal to $n = 21837$. Here, we focus on municipalities (second administrative units) as small areas and produce the MPI of the Pacific region using the Imp technique only. The variables used for the imputation and matching are the ones used in the simulation. Figure 3 shows the MPI estimates for municipalities in the Pacific region of Colombia.

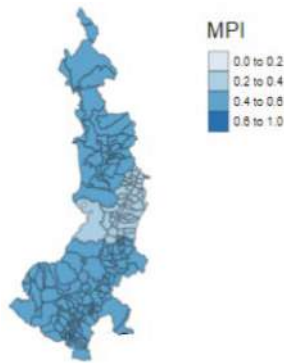


Fig. 3. Global MPI estimates for municipalities in the Pacific region of Colombia.

5 Conclusion

The article investigates two data integration methods for estimating poverty in small areas with missing indicators. The non-parametric two-step imputation approach, outperforms GLMM when model assumptions fail considerably. It produces less biased results and is less sensitive to model misspecification. Despite drawbacks such as variable selection complexity, both methods yield satisfactory quality for MPI estimates. Considering uncertainty from the estimated indicators is object of future research. Future research will explore spatial microsimulation as a potential method in order to deal with model misspecification issues.

References

1. Moretti, A., Shlomo, N., Sakshaug, J.W.: Small area estimation of latent economic well-being. *Sociol. Methods Res.* **50**(4), 1660–1693 (2021)
2. Joint Research Centre-European Commission. Handbook on constructing composite indicators: methodology and user guide. OECD publishing (2008)
3. Shlomo, N., et al.: The making of EVENS. In: Finney, N., Nazroo, J., Bécares, L., Kapadia, D., Shlomo, N. (eds.) *Racism and ethnic inequality in a time of crisis: findings from the evidence for equality national survey*, pp. 11–29. Policy Press (2023)
4. Arias-Salazar, A., Gutiérrez, A., Mancero, X., Guerrero-Gómez, S., Rojas-Perilla, N., Zhang, H.: Small area estimation for composite indicators: the case of multidimensional poverty incidence. arXiv preprint [arXiv:2304.03901](https://arxiv.org/abs/2304.03901) (2023)
5. Alkire, S., Foster, J.: Counting and multidimensional poverty measures. OPHI working paper 7 (2007)
6. D’Orazio, M., Di Zio, M., Scanu, M.: *Statistical Matching: Theory and Practice*. John Wiley & Sons (2006)
7. Im, J., Kim, J.K., Fuller, W. A. Two-phase sampling approach to fractional hot deck imputation. In *Proceedings of the Survey Research Methods Section*, pp. 1030–1043. (2015)
8. Chambers, R., Chandra, H., Salvati, N., Tzavidis, N.: Outlier robust small area estimation. *J. R. Stat. Soc. Ser. B Stat Methodol.* **76**(1), 47–69 (2014)



A Tool for Identifying Drivers Impacting Students' Career Aspirations

Flaminia Musella²(✉), Brunella Botte¹, and Giada Marinensi¹

¹ Link Campus University, Via del Casale di San Pio V 44, Roma, Italy
{b.botte,g.marinensi}@unilink.it

² Università degli Studi Roma Tre, Via del Castro Pretorio 20, 00185 Roma, Italy
flaminia.musella@uniroma3.it

Abstract. University identity, defined by students' sense of belonging to their respective higher institutions, serves as a form of social identification enabling students to boost their self-concept or self-image through affiliation with the *alma mater*. Considering the paradigm shift of students as customers and higher institutions as service providers, university identity offers a promising avenue for exploring how branding influence students' confidence and career aspirations. This study explores factors, beyond brand identity, impacting on students' perception to reach their educational goals. In this work, a scale was developed and evaluated; its psychometric properties assessed initially on a pilot sample, followed by a validation on a larger sample and via bootstrap samples for model robustness. By delving into these factors, this research provides insights into areas where institutions can enhance support for student success and well-being, particularly through innovative teaching methods.

Keywords: scale assessment · students expected professional outcomes · structural equation models

1 Introduction

Upon entering university, students possess a range of aspirations spanning learning, personal development, career objectives, and social connections (Pascarella & Terenzini, 2005; Sánchez-Sandoval & Verdugo, 2016). Despite brand identity support students' perception to reach their goals, other factors, both within and outside the university environment, influence students' career aspiration (Zepke et al., 2010). University factors, categorized as “hard” (e.g., courses, disciplines) and “soft” (e.g., physical space, empathy, management), play a significant role. This study focuses in identifying factors within the university environment that empower students in goal fulfillment and provides a tool for institutions to gauge their support for students and identify areas for enhancement. Conducted under the Empower Competencies for Online Learning in HE (ECOLHE) project, co-funded by the European Union's ERASMUS + program, the paper presents its methodology in Sect. 2, results in Sect. 3, and a brief discussion in Sect. 4.

2 Methodological Approach

The aim of this study is to develop a tool for assessing students' perceptions of universities' ability to empower their career aspirations. To reach the goal, a questionnaire was created, validated, and assessed, following a rigorous methodology including pilot studies and simulations. Drawing from established methodologies proposed by Trochim (2006) and Dimitrov (2012), as well as Furr (2011), the process was organized into three main phases: (i) scale creation, (ii) verification of its psychometric properties, and (iii) questionnaire assessment (see Fig. 1).

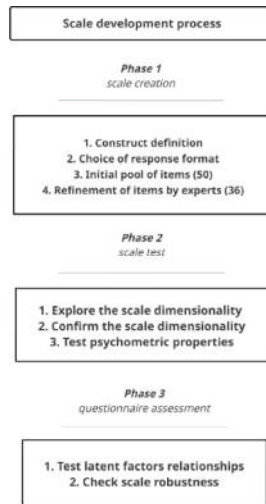


Fig. 1. The methodological process

3 Results

Phase one commenced with a literature review aimed at identifying 50 redundant items, which were subsequently refined by experts to 36. The selected response format was a 5-point Likert scale. Phase two was executed with a pilot sample comprising 505 Spanish students. The dimensionality of the scale was assessed through exploratory factor analysis (EFA) conducted on the first half (250 units) of the sample, revealing five factors and indicating the multidimensionality of the scale. Confirmatory factor analysis (CFA) was subsequently conducted on the remaining sample (255 units) to confirm the dimensionality of the scale. The resulting factors are:

- **F1: Teaching methods**, evaluating the innovation within teaching methodologies, learning approaches, as well as the tools and resources utilized in both educational activities and assessment methods.

- **F2: Educational environment and facilities** seeks to provide a comprehensive overview that combines the social aspects, such as the staff-student relationship, with the infrastructure, notably the digital infrastructure, within educational institutions.
- **F3: Teachers’ empathy and care**, focusing on aspects related to the teaching process and the relationship between students and faculty.
- **F4: Students’ commitment to the institution**, capturing students’ overall satisfaction with the learning experience, including how it aligns with their initial expectations.
- **F5: Students’ expected professional outcomes**, investigating their perceptions of how attending the university would impact their career aspirations empowerment.

Table 1 shows the status of CFA indicating that all loadings exceed the acceptable threshold (>0.60), signifying the relevance of items to their respective dimensions. Additionally, all p-values are statistically significant (<0.05), and the R^2 values exceed 0.4, indicating a strong and meaningful correlation between items and their assigned dimensions.

Table 1. Status of CFA

Fac	α	CR	AVE	Item	Est	S.E.	z-val	p-val	St. Load	R^2
F1	.79	.83	.57	To foster students’ learning, in the classes, the teachers:						
				use game elements or educational games	1.000				.734	.539
				use case studies	1.049	.081	12.920	.000	.770	.594
				use lab experiments and simulations	1.051	.075	14.071	.000	.772	.596
				assess students’ prior knowledge to orient personalised learning	1.004	.069	14.488	.000	.737	.543
F2	.85	.89	.70	The faculty organisation/structure is clear to me	1.000				.836	.700
				Announcements from the administrative staff are clear	.947	.037	25.806	.000	.792	.628
				Technology and learning portals (<i>e.g.</i> Moodle, Learning Management System) are effectively used	1.067	.037	29.120	.000	.893	.797

(continued)

Table 1. (continued)

Fac	α	CR	AVE	Item	Est	S.E.	z-val	p-val	St. Load	R ²
				ICT Tools and platforms are intuitively used	.987	.036	27.442	.000	.826	.682
F3	.91	.95	.87	Teaching staff is empathic	1.000				.919	.845
				Teaching staff provide students with the support they need	1.055	.025	42.528	.000	.969	.940
				Teachers are engaged in the teaching process	.995	.021	46.345	.000	.914	.836
F4	.89	.94	.85	I believe that studying at this University:						
				Matches my learning expectations	1.000				.960	.923
				Is really enjoyable	.948	.019	49.085	.000	.910	.829
				Is developing my soft skills	.928	.022	43.020	.000	.891	.794
F5	.84	.89	.74	I believe that studying at this University:						
				Will help me in acquiring a job or career-related knowledge and skills	1.000				.819	.671
				Will help me develop my critical thinking	1.078	.045	23.793	.000	.883	.779
				Will help me in team working	1.063	.047	22.652	.000	.870	.758

The psychometric properties of the scale were evaluated using the same subset of data, with the results summarized in Table 1. Firstly, we assessed the reliability of the scale. *Internal consistency*, examined through Cronbach's alpha, initially indicated slight redundancy ($\alpha = 0.97$) in the questionnaire. However, after refining the scale to five dimensions and 17 items, coherence was achieved, with an overall α equals 0.94 and individual dimension alphas ranging from 0.79 to 0.91. *Composite reliability*, indicating the extent to which factors are related to each other, was confirmed to be above the threshold (0.70) using the CR index.

Moving to validity, *convergent validity*—measuring the sensitivity of items to correlate with the latent construct—was assessed using the average variance extracted (AVE) index, which consistently exceeded 0.5, indicating good results. *Discriminant validity*, demonstrating the distinction between constructs, was confirmed when the correlation between constructs was less than the square root of AVE.

In phase three, the relationships between factors were examined using a structural equation model (SEM). The model, depicted in Fig. 2, represents latent factors as ellipses, relationships as arrows, and the strength of connections as coefficients. All relationships in the model were statistically significant ($p\text{-value} < 0.05$) and demonstrated relevant impacts.

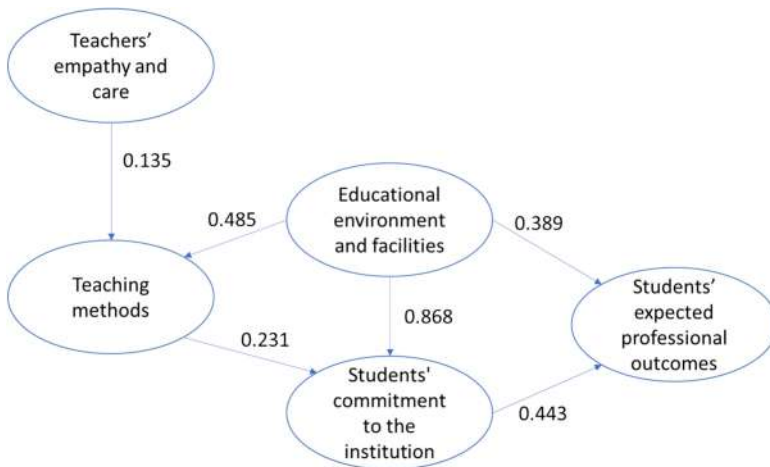


Fig. 2. Structural model

Furthermore, the goodness-of-fit (GOF) statistics of the model were highly promising, with a Goodness-of-Fit Index (GFI) of 0.995, Comparative Fit Index (CFI) of 0.998, Root Mean Square Error of Approximation (RMSEA) with a 90% confidence interval of 0.046 (ranging from 0.032 to 0.059), Standardized Root Mean Square Residual (SRMR) of 0.048, and Tucker-Lewis Index (TLI) of 0.998.

4 Discussion

The paper reveals several significant findings. *Students' expected professional outcomes* are directly and positively influenced by their *commitment to the institution* (0.443). *Educational environment and facilities* have a substantial impact on the target variable both directly and indirectly, with a total effect of 0.773, indicating their importance for achieving the desired outcomes. Moreover, *teaching methods* are affected by the *educational environment and facilities* (0.485) and to a lesser extent by *teachers' empathy and care* (0.135). Conversely, *students' commitment to the institution* is strongly influenced by the *educational environment and facilities*, both directly and indirectly, with a total

effect of 0.716. Additionally, it is somewhat dependent on teaching methods (0.231), suggesting that student engagement is more influenced by the environment than the educational methods themselves.

To enhance the confidence in the psychometric properties, the questionnaire was administered to a larger sample consisting of 227 Italian respondents. The psychometric properties were re-evaluated on this new dataset, and the structural integrity of the model was confirmed. Additionally, the goodness-of-fit (GOF) of the model was computed on 1000 bootstrap samples, as recommended by prominent literature (Cheung & Lau, 2008). The average performance of the model across these samples indicated a good overall fit, further solidifying the robustness of the findings.

References

- Dimitrov, D.M.: Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields. American Counseling Association, Alexandria, VA (2012)
- Furr, R.M.: Scale Construction and Psychometrics for Social and Personality Psychology. Sage Publications, New Delhi, IN (2011). <https://doi.org/10.4135/9781446287866>
- Pascarella, E.T., Terenzini, P.T.: How College Affects Students: A Third Decade of Research, vol. 2. Jossey-Bass, An Imprint of Wiley. 10475 Crosspoint Blvd, Indianapolis, IN 46256 (2005)
- Sánchez-Sandoval, Y., Verdugo, L.: Desarrollo y validación de la Escala de Expectativas de Futuro en la Adolescencia (EEFA). *Anales de Psicología* **32**(2), 545 (2016)
- Trochim, W.M.K.: Research Methods Knowledge Base (2006). www.socialresearchmethods.net/kb/index.php. Accessed 14 Sep 2007
- Zepke, N., Leach, L., Butler, P.: Non-institutional influences and student perceptions of success. *Stud. High. Educ.* **36**(2), 227–242 (2011)



Changes in Seasonal Concentration of Tourist Flows: The Role of Tourism Market Segments

Mauro Mussini^(✉)

Department of Economics, Management and Statistics, University of Milano-Bicocca,
Piazza dell'Ateneo Nuovo, 1 - 20126 Milan, Italy
mauro.mussini1@unimib.it

Abstract. Seasonal concentration in tourist flows is often seen as an undesirable consequence of the tourism industry growth in a tourist destination because of its negative impacts on environment, use of natural resources and infrastructures, quality of life for inhabitants. Local governments can take various initiatives to counter seasonality in tourist flows, such as market diversification or different pricing strategies for distinct tourist segments. To assess the effectiveness of such initiatives, a statistical toolkit for measuring the changes in seasonal concentration is needed. This paper suggests a decomposition that separates the roles played by different segments of tourists in changing the pattern and magnitude of seasonal concentration in tourist flows.

Keywords: Concentration · Seasonality · Tourism

1 Introduction

A huge concentration of tourist flows in relatively short periods of the year often implies several disadvantages for a tourism destination. Such a concentration of tourists in a few months or weeks may imply an overuse of natural resources and infrastructures in the high peak season, an increase in pollution and prices, a deterioration of the quality of life for inhabitants and of the touristic experience for the visitors themselves. The local governments of tourist destinations often take initiatives to counter seasonal concentration of tourist flows. In that scenario, monitoring the changes in seasonal concentration is crucial for evaluating the impact of such strategies to counter seasonality in tourist flows.

Several statistical indicators have been used in the literature on tourism to measure seasonal concentration of tourist flows [1, 2]. However, most of these indicators are suitable for measuring seasonal concentration in a certain year but not to track changes in seasonal concentration as they neglect the changes that may occur in the seasonal pattern, i.e. the way seasonal fluctuations reoccur at the same times every year. Recently, Grossi and Mussini [3] have proposed a decomposition of the change in seasonal concentration which separates a component due to the change in the seasonal pattern, i.e. a seasonal stability component, from a component measuring the change in the seasonal magnitude, i.e. a pure seasonal magnitude component. This paper shows that the seasonal stability

and magnitude components can be split into the contributions that come from different segments of the tourism market (e.g. residents and non-residents; retirees and non-retirees). This additional information is quite relevant to policymakers in the tourism sector as they can know the role played by each segment of tourists in changing the seasonal concentration both in terms of stability and magnitude. For instance, if the tourist flows of a certain segment (e.g. retirees) became more evenly distributed throughout the year for effect of some incentives, the impact of such a change in distribution on seasonal stability and magnitude can be measured.

2 The Components of the Change in Seasonal Concentration

Let Y stand for a variable that represents tourist arrivals in a tourist destination and $y_{1,t}, \dots, y_{n,t}$ be the n observations of Y_t in the n periods into which the year t is divided (n is equal to 12 in case of monthly observations). The Gini index measuring the concentration of tourist arrivals in year t is [4].

$$G_t = \frac{2cov[Y_t, r(Y_t)]}{n\bar{y}_t}, \quad (1)$$

where $\bar{y}_t$ is the average number of tourist arrivals per period in year t and $r(Y_t)$ is the rank of the observation of Y_t in a certain period of the year.

Now suppose that the distribution of tourist arrivals is observed in year $t + k$. The change in the seasonal concentration from year t to $t + k$ is measured by the difference between the Gini index in $t + k$ and that in t :

$$\Delta G = G_{t+k} - G_t = \frac{2cov[Y_{t+k}, r(Y_{t+k})]}{n\bar{y}_{t+k}} - \frac{2cov[Y_t, r(Y_t)]}{n\bar{y}_t}. \quad (2)$$

ΔG does not necessarily capture all aspects of the change in seasonality. For instance, suppose that the monthly observations of tourist arrivals in year t are the same in magnitude of those in year $t + k$ but that they occur in different months, i.e. fluctuations do not reoccur in the same months every year. In such a situation, the Gini index in $t + k$ is equal to the Gini index in t and ΔG is null, suggesting that seasonality has remained unchanged since t although the distribution of arrivals across months has changed. To track this change in seasonality, a measure of seasonal stability is needed. Grossi and Mussini [3] obtained such a measure of seasonal stability by decomposing ΔG .

Let $C_{t+k|t}$ be the concentration coefficient of tourist arrivals in year $t + k$ obtained by sorting the observations in the n periods according to the ranking of periods by arrivals in year t instead of that in $t + k$:

$$C_{t+k|t} = \frac{2cov[Y_{t+k}, r(Y_t)]}{n\bar{y}_{t+k}}. \quad (3)$$

By adding and subtracting $C_{t+k|t}$ to the right-hand side of Eq. (2), the difference in the Gini index is broken down into two components:

$$\begin{aligned}\Delta G &= \left\{ \frac{2\text{cov}[Y_{t+k}, r(Y_{t+k})]}{n\bar{y}_{t+k}} - \frac{2\text{cov}[Y_{t+k}, r(Y_t)]}{n\bar{y}_{t+k}} \right\} \\ &\quad + \left\{ \frac{2\text{cov}[Y_{t+k}, r(Y_t)]}{n\bar{y}_{t+k}} - \frac{2\text{cov}[Y_t, r(Y_t)]}{n\bar{y}_t} \right\} \\ \Delta G &= R + M.\end{aligned}\tag{4}$$

R in Eq. (4) is the re-ranking component measuring the change in concentration due the changes in the ranking of periods by tourist arrivals from t to $t + k$. R is null if the ranking of periods in $t + k$ is the same as that in t , whereas it equals $2 G_{t+k}$ if the ranking of periods in year $t + k$ is exactly reversed when compared to the ranking of periods in year t . As R tracks every deviation in the seasonal pattern, this component can be seen as a measure of seasonal stability. The larger the re-ranking component, the more unstable the seasonal pattern.

M in Eq. (4) is the magnitude component measuring how much tourist arrivals in year $t + k$ are more (or less) concentrated in the periods that are the peak periods in year t . Since M is calculated by keeping periods sorted according to their ranking in t , the seasonal pattern is assumed to be the same in both years. The magnitude component is positive (negative) when the magnitude of seasonality has increased (decreased) with respect to the initial distribution of arrivals among periods.

3 A Decomposition by Tourism Market Segmentation

Components R and M can be broken down into the contributions coming from different segments of tourists. Suppose that Y_t is given by the sum of arrivals from different tourist segments (e.g. retirees and non-retirees) and that we have m tourist segments, so that $Y_t = X_{1,t} + \dots + X_{j,t} + \dots + X_{m,t}$. Lerman and Yitzhaki [4] suggested a decomposition of the Gini index by using the Gini correlation. The Gini correlation between variables $X_{j,t}$ and Y_t is defined as

$$\Gamma(X_{j,t}, Y_t) = \frac{\text{cov}(X_{j,t}, r(Y_t))}{\text{cov}(X_{j,t}, r(X_{j,t}))}.\tag{5}$$

$\Gamma(X_{j,t}, Y_t)$ ranges from -1 to 1, and the Gini index of Y_t can be written as

$$G(Y_t) = \sum_{j=1}^m S_j G(X_{j,t}) \Gamma(X_{j,t}, Y_t),\tag{6}$$

where $S_j = n\bar{x}_{j,t}/n\bar{y}_t$ is the share of $X_{j,t}$ in Y_t , and $G(X_{j,t})$ is the Gini index of $X_{j,t}$.

Given Eqs. (5) and (6), the seasonal stability component can be split by segment:

$$\begin{aligned}
 R &= \frac{2cov[Y_{t+k}, r(Y_{t+k})]}{n\bar{y}_{t+k}} - \frac{2cov[Y_{t+k}, r(Y_t)]}{n\bar{y}_{t+k}} \\
 R &= \sum_{j=1}^m \frac{n\bar{x}_{j,t+k}}{n\bar{y}_{t+k}} \frac{2cov(X_{j,t+k}, r(X_{j,t+k}))}{n\bar{x}_{j,t+k}} \frac{cov(X_{j,t+k}, r(Y_{t+k}))}{cov(X_{j,t+k}, r(X_{j,t+k}))} - \\
 &\quad \sum_{j=1}^m \frac{n\bar{x}_{j,t+k}}{n\bar{y}_{t+k}} \frac{2cov(X_{j,t+k}, r(X_{j,t+k}))}{n\bar{x}_{j,t+k}} \frac{cov(X_{j,t+k}, r(Y_t))}{cov(X_{j,t+k}, r(X_{j,t+k}))} \\
 R &= \sum_{j=1}^m S_{j,t+k} G(X_{j,t+k}) \Gamma(X_{j,t+k}, Y_{t+k}) - \sum_{j=1}^m S_{j,t+k} G(X_{j,t+k}) \Gamma(X_{j,t+k}, Y_t) \\
 R &= \sum_{j=1}^m S_{j,t+k} G(X_{j,t+k}) [\Gamma(X_{j,t+k}, Y_{t+k}) - \Gamma(X_{j,t+k}, Y_t)] \\
 R &= \sum_{j=1}^m R_j.
 \end{aligned} \tag{7}$$

where $R_j = S_{j,t+k} G(X_{j,t+k}) [\Gamma(X_{j,t+k}, Y_{t+k}) - \Gamma(X_{j,t+k}, Y_t)]$ is the contribution of segment j to the seasonal stability component. It is worth mentioning that the sign of R_j depends only on the difference between the Gini correlations $\Gamma(X_{j,t+k}, Y_{t+k})$ and $\Gamma(X_{j,t+k}, Y_t)$ for each $X_{j,t+k}$, with $j = 1, \dots, m$. While the seasonal stability component R is always non-negative (see Sect. 2), the contribution to R from a single tourist segment can be negative. This occurs when $X_{j,t+k}$ is more Gini-correlated to Y_t than to Y_{t+k} , i.e. the distribution of arrivals from the tourist segment j among periods in year $t+k$ is more consistent with the seasonal pattern in year t . Thus, a negative value of R_j means that the tourist segment j contributes to the stability of the seasonal pattern over time (remember that the larger R , the more unstable the seasonal pattern; see Sect. 2).

The decomposition of the seasonal magnitude component is

$$\begin{aligned}
 M &= \frac{2cov[Y_{t+k}, r(Y_t)]}{n\bar{y}_{t+k}} - \frac{2cov[Y_t, r(Y_t)]}{n\bar{y}_t} \\
 M &= \sum_{j=1}^m S_{j,t+k} G(X_{j,t+k}) \Gamma(X_{j,t+k}, Y_t) - \sum_{j=1}^m S_{j,t} G(X_{j,t}) \Gamma(X_{j,t}, Y_t) \\
 M &= \sum_{j=1}^m M_j
 \end{aligned} \tag{8}$$

where $M_j = S_{j,t+k} G(X_{j,t+k}) \Gamma(X_{j,t+k}, Y_t) - S_{j,t} G(X_{j,t}) \Gamma(X_{j,t}, Y_t)$. M_j is the contribution to the seasonal magnitude component that comes from the tourist segment j . The interpretation of M_j is in line with that of M : a positive (negative) value of M_j means that the tourist segment j contributes to increase (decrease) the magnitude of seasonality in the overall distribution of tourist arrivals among periods.

4 A Numerical Example

Suppose that monthly tourist arrivals in a certain destination can be divided into two segments by tourist type: retirees and non-retirees. In addition, suppose that such a tourist destination was affected by a high seasonal concentration in arrivals in year t and that local policymakers set up a pricing strategy addressed to retirees with the aim of reducing seasonality between years t and $t + k$. Figure 1 shows the decomposition of the seasonal stability and magnitude components by segment. The contribution to R

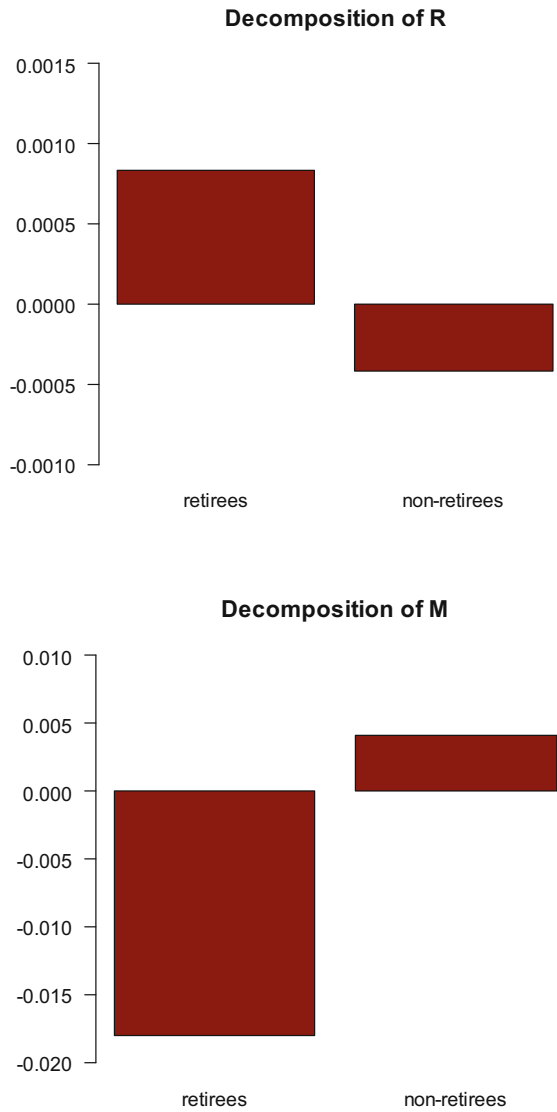


Fig. 1. Decomposition of the seasonal stability and magnitude components by tourist segment.

from the retiree segment is positive (0.00083), suggesting that there was a change in the ranking of months by retiree tourist arrivals from year t to $t + k$ and the change in the relative distribution of retiree tourists across months weakened the initial seasonal pattern. The contribution to M from retirees is negative (-0.01801), suggesting that the change in the relative distribution of retiree tourists decreased the magnitude of seasonal concentration. Such results indicate that adopting a specific pricing strategy for retirees was effective in mitigating seasonal concentration, although the impact of this initiative on the change in seasonal concentration was small (-0.01718). The contribution to R from the non-retiree segment is negative (-0.00042), suggesting that the non-retiree tourist arrivals spread out across months of year $t + k$ in a way that preserved the seasonal pattern. The contribution to M from non-retirees is positive (0.00406), indicating that the change in the relative distribution of non-retiree tourists increased the magnitude of seasonal concentration.

5 Conclusion

The seasonal concentration of tourist flows is generally measured by considering the number of tourist arrivals at the aggregate level and using a set of statistical indicators. This way of examining seasonal concentration neglects the roles that the different segments of the tourism market play in determining the degree of seasonal concentration or the changes in seasonal concentration. This paper further develops the Grossi and Mussini [3] decomposition of the change in seasonal concentration to separate the contributions from different tourist segments to the stability and magnitude of seasonality in tourist arrivals.

The next step of the research is the application of the decomposition to real data on arrivals in a tourist destination where seasonal concentration is high and represents a phenomenon that the local government tries to mitigate. The results of the decomposition could provide policymakers with additional information, which could be used to design specific strategies to counter seasonality on the basis of the role of each tourist segment in determining changes in seasonal concentration.

References

1. Fernández-Morales, A.: Decomposing seasonal concentration. *Ann. Tour. Res.* **30**, 942–956 (2003)
2. Fernández-Morales, A., Mayorga-Toledano, M.C.: Seasonal concentration of the hotel demand in Costa del Sol: a decomposition by nationalities. *Tour. Manage.* **29**, 940–949 (2008)
3. Grossi, L., Mussini, M.: Seasonality in tourist flows: decomposing and testing changes in seasonal concentration. *Tour. Manage.* **84**, 104289 (2021)
4. Lerman, R.I., Yitzhaki, S.: Income inequality effects by income source: a new approach and applications to the United States. *Rev. Econ. Stat.* **67**(1), 151–156 (1985). <https://doi.org/10.2307/1928447>



Evaluating Volleyball Setters Using Product Partition Models

Martina Narcisi¹(✉), Garritt L. Page², and Gilbert W. Fellingham²

¹ Department of Statistics, University of Bologna, Bologna, Italy
martina.narcisi2@unibo.it

² Department of Statistics, Brigham Young University, Provo, USA
page@stat.byu.edu gwf@byu.edu

Abstract. Assessing volleyball player performance has long been a subject of interest in literature, with a focus on skills evaluation. In particular, the evaluation of setters presents unique challenges due to their neither defensive nor offensive role, despite their central importance as playmakers. Setters influence team success in complex ways, making it difficult to isolate individual contributions. Traditional assessment methods heavily rely on attack quality, which may not accurately reflect setter performance. To address this, we propose a novel approach that considers the precise location of sets before the attack that can kill or not. By categorizing sets based on location and outcome, we aim to develop a metric that accounts for setter influence beyond offensive execution. Our methodology promises to identify top setters in the 2018 men's Championship, offering valuable insights into this crucial yet understudied aspect of volleyball performance evaluation.

Keywords: Product partition model · spatial clustering · evaluation performance · volleyball

1 Introduction

Like most team sports, success in volleyball depends heavily on the way in which the skill set of individual players interact towards contributing to overall team effectiveness. As such, skill assessment in volleyball has a presence in the sports analytics literature (Florence et al., 2008; Silva et al., 2014; Drikos et al., 2019; Fellingham 2022). Typically, player performance is evaluated by assigning values on a 5-point scale to various skills. Crucially, player performance rating depends not only on the quality of the skill performed, but how the opposing team handles it. For example, it is common to rate the effectiveness of a serve by evaluating the quality of its reception. Thus, a good serve can be rated poorly if the receiver is able to handle it well while a poor serve may be evaluated as effective if the receiver poorly executes a pass. Thus, assessing volleyball skills, especially when evaluating setters, poses a significant challenge due to potential confounding factors. Additionally, the setter position is widely regarded as one of the most critical in volleyball, highlighting the need for good evaluation metrics.

Evaluating a volleyball setter’s performance during a match has been notoriously challenging. This is because it is not straightforward to disentangle team success from that of the individual contributors who interact in complex ways. This is particularly true when assigning a performance score to the setters, who do not hold an offensive position but rather act as the playmaker of the game. They are the team’s backbone, managing them play, yet seldom executing the final offensive move. In general, the team’s success in a set relies heavily on the quality of the attack, a method which places significant emphasis on the proficiency level of the player executing the offensive action. Given that this dependency is overly tied to the attacker’s performance, in this study, we have opted to weigh this information by considering the precise location in which the setter sets the ball. Although players must start in fixed locations before the serve, once the ball is put in play, players generally move to other locations on the court. In particular, the setter always goes near the red cross in Fig. 1. From that position, he moves wherever necessary, traversing the entire court as required. Our strategy involves initially grouping sets according to their location and outcome. Subsequently, utilizing this clustering, we devise a metric that assigns a value corresponding to the quality of the set, factoring in its location. Finally, this methodology will enable us to ascertain the best setters in the 2018 men’s World Championship.

2 The Data and the Model

We use data from the 2018 men’s World Championship that was collected by HUDL (a U.S. based company providing statistical and video review tools for teams at all levels of American football, baseball, volleyball, soccer, and lacrosse). The data contains information about every touch by every player in the tournament and a skill evaluation of each touch. We collected data regarding the positions of both setters and hitters who made contact with the ball during each team’s possession. As already mentioned, we seek to evaluate the setter’s performance using his exact spatial location when he set and the outcome of the hitter that follows.

For each match of the men’s World Championship volleyball, the location on the court at which each set took place was recorded. The result of the set (kill or not) was also recorded. We will denote the result of the i th set for the j th setter as $Y_{ij} \in \{0, 1\}$ where 1 corresponds to a kill. Additionally, we will use $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij})$ denote the coordinates of the court at which the i th set for the j th setter took place.

Our strategy involves initially grouping sets by their location and outcome. Subsequently, utilizing this clustering, we devise a metric that evaluates the quality of the set, considering its location. Ultimately, we aim to identify the superior setters in the championship.

The response we are considering is binary in that the set results in a kill or not. To model these data and produce a grouping that is based on the location of the set and the result, we will use the ordinal Product Partition Model (PPMx)

procedure that was detailed in? and with accompanying software in Page and J. Quinlan (2023).

Since the response is binary, it is natural to consider latent variable ordinal data models as those described in Bao and Hanson (2015) and in Kottas et al. (2005). For the sake of completeness, we detail the model in its entirety. Let $Y_i \in \{0, 1\}$ be the binary response measured for the i th set for $i = 1, \dots, n$, $n = 12496$. Further, let $\mathbf{X}_i = (X_{i1}, X_{i2})$ denote longitude and latitude at which the i th set was executed. As is done in Bao and Hanson (2015), real-valued latent scores $Z_1, \dots, Z_n$ are introduced such that for $-\infty = \gamma_0 < \gamma_1 < \gamma_2 = \infty$,

$$Y_i = \ell \Leftrightarrow \gamma_\ell < Z_i \leq \gamma_{\ell+1} \text{ for } \ell = 0, 1.$$

The appeal of employing the methods described in Bao and Hanson (2015) and Kottas et al. (2005) is that the values selected for γ are immaterial, so long as the model is flexible enough to assign sufficient probability mass to each $(\gamma_\ell, \gamma_{\ell+1}]$ interval. Since assigning a PPMx prior to ρ_m results in modelling the latent ordinal scores with a mixture (which is essentially the same data model arising from marginalizing the random measure in the approach by Bao and Hanson (2015)), we conclude that the model we specify affords sufficient flexibility and, therefore, the γ can be selected arbitrarily. In light of this, we set $\gamma_1 = 0$. The remainder of the model is expressed hierarchically after introducing latent cluster labels $s_1, \dots, s_n$:

$$\begin{aligned} Z_i \mid \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}, s_i &\sim N(\mu_{s_i}^*, \sigma_{s_i}^{2*}), \\ (\mu_j^*, \sigma_j^*) &\sim N(\mu_0, \sigma_0^2) \times Unif(0, A), \\ Pr(\rho \mid \mathbf{X}) &\propto \prod_{j=1}^k c(S_j)g(\mathbf{X}_j^*), \end{aligned} \tag{1}$$

where $c(S_j)$ and $g(\mathbf{X}_j^*)$ are set to functions. Further, we assume $\mu_0 \sim N(0, 10^2)$, $\sigma_0 \sim Unif(0, 10)$, and set $A = 0.1$.

As a first step, we model fit an ordinal PPMx to all the sets, Fig. 1 shows the plot in which it is possible to observe the grouping of sets based on their location, without distinction among different setters. We also get the overall probabilities, $p_{j,i}^{over}$, related to each unit j expressing the probability that the j -attack can kill considering that it belongs to cluster $i = 1, \dots, 6$.

After modeling the ordinal PPMx on the overall sets, we now apply it separately for each setter $s = 1, \dots, 44$ obtaining the probabilities associated with the different setters, $p_{j,i}^s$ related to each setter s . They express the probability that the j th-attack, set by setter s , results in a kill given that it belongs to cluster $i = 1, \dots, 6$.

We suggest the following index, the *Setter score* (SS), capable of evaluating the performance of each setter:

$$SS_s = \sum_{i=1}^6 \sum_{j=1}^{n_i} \frac{p_{j,i}^{over} - p_{j,i}^s}{n_i} w_i \tag{2}$$



Fig. 1. Cluster assignment of sets based on their location and hitter's result. The red cross represents the optimal set position.

where

$$w_i = \left(\frac{1}{n_i} \sum_{j=1}^{n_i} p_{j,i}^{over} \right)^{-1}$$

is the weight related to each cluster i in the ordinal PPMx on the overall sets and n_i is the number of sets in each cluster. This means that the higher the setter score, the better the performance of the setter is evaluated, namely

$$>> SS_s \iff \text{better performance of setter } s.$$

The histogram in Fig. 2 shows the empirical distribution of the SS variable. The empirical curve depicted on the graph cannot be assimilated to a known curve. This is because the dataset did not allow for the analysis of a sufficiently large number of setters s . We are particularly interested in identifying the individuals that characterize the right tail of this distribution. They will be candidates to be the best setters of the championship under consideration.

3 Conclusion

We have shown a methodology that allows setters to be evaluated based on the position in which they set and the result of the attack after it. It may appear that the evaluation of the preceding reception has been overlooked, namely the

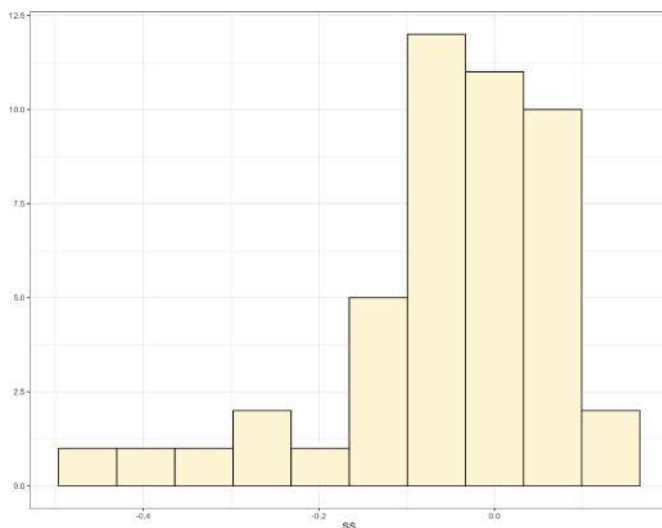


Fig. 2. Empirical distribution of the setter score.

first of the three touches, but in reality, this information is correlated with the setter's positioning when touching the ball. In this paper, it becomes evident the aim is to evaluate the setter's performance, considering how a successful or unsuccessful reception can affect their work and lead them to move more or less from the optimal setting point (red cross in Fig. 1). Additionally, the outcome of the attack is taken into account: if the attack kills, it is reasonable to assume that the set was effective. The highest setter score is assigned to the setter who collects the most winning attacks following their set, especially when made from a non-optimal position.

References

- Bao, J., Hanson, T.E.: Bayesian nonparametric multivariate ordinal regression. *Can. J. Stat.* **43**, 337–357 (2015)
- Drikos, S., Ntzoufras, I., Apostolidis, N.: Bayesian analysis of skills importance in world champions men's volleyball across ages. *Int. J. Comput. Sci. Sport* **18**, 24–44 (2019). <https://api.semanticscholar.org/CorpusID:198181418>
- Fellingham, G.W.: Evaluating the performance of elite level volleyball players. *J. Quant. Anal. Sports* **18**(1), 15–34 (2022). <https://doi.org/10.1515/jqas-2021-0056>
- Florence, L.W., Fellingham, G.W., Vehrs, P.R. and Mortensen, N.P.: Skill evaluation in women's volleyball. *J. Quant. Anal. Sports* **4**(2). <https://doi.org/10.2202/1559-0410.1105>
- Kottas, A., Müller, P., Quintana, F.: Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Stat.* **14**, 610–625 (2005)
- Page, G.L., Quinlan, J.: ppmSuite: a collection of models that employ product partition distributions as a prior on partitions. R package version 0.3.3 (2023). <https://CRAN.R-project.org/package=ppmSuite>

- Page, G.L., Quintana, F.A., Rosner, G.L.: Discovering interactions using covariate informed random partition models. *Ann. Appl. Statist.* **15**(1), 1–21 (2021). <https://doi.org/10.1214/20-AOAS1372>
- Silva, M., et al.: Game-related volleyball skills that influence victory. *J. Hum. Kinet.* **41**, 173–179 (2014). <https://api.semanticscholar.org/CorpusID:98917>



Combining Price Indexes and Fuzzy Clustering to Measure Satisfaction in Niche Tourism Destinations

Consuelo Rubina Nava^{1(✉)}, Tiziana Ciano¹, and Massimiliano Ferrara²

¹ Department of Economics and Political Science, University of Aosta Valley, Aosta, Italy

{c.nava,t.ciano}@univda.it

² Department of Law, Economics and Human Sciences, University Mediterranea of Reggio Calabria, Reggio Calabria, Italy

massimiliano.ferrara@unirc.it

<https://www.univda.it/docenti>,

<http://www.decisionslab.unirc.it/it/massimiliano-ferrara/>

Abstract. Niche tourism is an expanding global industry with unique and customized experiences catering to specific market segments. Accommodation facilities are pivotal in accommodating and catering to niche tourists, providing personalized and tailored experiences. A two-step procedure is proposed to compare satisfaction levels across accommodation options based on Booking.com service scores. This approach relies on i) the Country Product Dummy index and ii) a Fuzzy clustering method for indirect relational data. As a case study for this methodology, accommodation facilities in the Lower Aosta Valley are examined, serving as a notable example of a niche tourism destination in Italy.

Keywords: Country product dummy index · Gower's distance · fuzzy clustering · niche tourism

1 Introduction

In recent years, the tourism industry has increasingly valorized the role of customer experience and feedback [1]. This has been possible thanks to the advent of big data, which has allowed for extensive integration of information and communication. The development of online platforms, such as Booking.com, is an essential source of information and a tool for evaluating the performance of the various tourist accommodation facilities [3]. Thus, reviews and accommodation scores have emerged as a crucial determinant in the choice process. This becomes even more important in niche tourism destinations [2]. The most common forms of niche tourism are food and wine tourism, cultural tourism, adventure tourism, and ecological tourism, to name a few. Instead of mass tourism, niche tourism provides travelers a more authentic experience aligned with their passions. It benefits local economies through sustainable opportunities, preserves cultures,

and promotes less-explored destinations. Additionally, it showcases a commitment to environmental and social responsibility, attracting conscious travelers. In Italy, the small-scale endeavor of Lower Aosta Valley could be considered a niche tourism destination given that it offers an opportunity for slow tourism, thematic hiking trails, a very mild climate, and typical/unique local food and wine offerings. Tourism has a powerful economic impact in this territory and represents one of the main economic activities, so local development policies are usually tourism-oriented. However, as a niche tourism destination, the Lower Aosta Valley faces many challenges: higher unit costs, limited capacity for revenue generation, inflexibility to market changes, over-reliance on specific markets, and potential environmental impacts [2]. For local accommodation facilities, acknowledging these drawbacks and evaluating/comparing their review scores remains a crucial aspect [3].

To this aim, this contribution builds a synthetic indicator of the different accommodation facilities located in the Lower Aosta Valley and registered on Booking.com. This indicator, based on the review scores collected by the booking platform over different satisfaction dimensions (staff, services, cleaning, comfort, quality-price ratio, location, free Wi-Fi), i) must represent a comparative tool between accommodation facilities and ii) must serve as a criterion for the supply segmentation. To attain these two purposes, a two-step methodological proposal is introduced. The first step measures variations in the tourist score evaluation of the distinct accommodation facilities via the Country Product Dummy (CPD hereafter) index [4] to fulfill point i). The second step groups accommodation facilities according to the computed CPD index and other dichotomous/categorical variables to fulfill point ii). The clusterization rests on a Fuzzy clustering method for indirect, relational data [5] based on Gower's distance [6]. This two-step methodological proposal overcomes the general shortcomings of descriptive statistics [7] in analyzing the scores assigned to the different accommodation facilities in the territory of interest.

The dimensionless and comparable CPD index here represents variations in review scores. It is used in a multilateral perspective (i.e., to compare different accommodation facilities given the selection of a base one). The CPD index becomes a more comprehensive tool for synthesizing evaluations of the different services by tourists. A positive variation in the index indicates increased satisfaction levels compared to the base accommodation facility: the more significant the variation, the higher the satisfaction. In this context, the CPD index is computed based on the scores assigned to different evaluation criteria included in the reference basket (staff, services, cleaning, comfort, quality-price ratio, location, free Wi-Fi) for accommodation facilities. Thus, the CPD index measures the variation in satisfaction levels across accommodation facilities.

The remainder of the paper is structured as follows. Section 2 explains the two steps methodology while Sect. 3 presents the data employed as a test-bed for the methodological proposal and the related results. Section 4 concludes the paper envisaging further research and highlighting the main limitations.

2 Methodology

Step 1: Country Product Dummy index. The first step estimates the CPD index [4] within the stochastic approach to analyze the variations of the scores assigned to each accommodation facility. Let the score of the i^{th} service ($i = 1, \dots, I$) for the j^{th} accommodation facility ($j = 1, \dots, J$) be denoted with p_{ij} , while PPP_j be the score equivalent of the purchasing power parity (PPP) of the j^{th} accommodation compared to a reference one (J). Accordingly, $PPP_{jk} = PPP_k / PPP_j$ is the PPP between currencies of any two accommodation facilities j and k , with $j, k = 1, \dots, J$. Given the Jevons geometric specification, the additive form of the model yields the estimation of the CPD index

$$\ln p_{ij} = \ln p_i + \ln PPP_j + \ln u_{ij}^* = \sum_{i=1}^I \eta_i D_i^P + \sum_{j=1}^{J-1} \pi_j D_j^C + \varepsilon_{ij} \quad (1)$$

where D_i^P takes values 1 for service i and 0 otherwise, D_j^C is equal to 1 for accommodation j and 0 otherwise, and ε_{ij} is the random error term. This model can be estimated by OLS obtaining:

$$\hat{\pi}_j = \frac{1}{N} \sum_{i=1}^N (\ln p_{ij} - \ln p_{iM}) \rightarrow P\hat{P}P_j = e^{\hat{\pi}_j}.$$

Step2: Fuzzy Clustering. The second step clusters accommodation facilities considering the estimated CPD index and other dichotomous/categorical variables such as sustainability levels, digitalization, type of accommodation facility, and municipality/localization. Given the presence of mixed variables, a Fuzzy clustering approach for relational data [5] using Gower's distance [6] is a suitable approach. The non-Euclidean fuzzy relational data clustering algorithm (NEFRC) is selected. NEFRC rests on the following minimization problem, which allows its use with all kinds of dissimilarity measures $d(\mathbf{x}_i, \mathbf{x}_j)$ computed between observations i and j , given r observed variables:

$$\min_{\mathbf{U}} \sum_{g=1}^k \frac{\sum_{i=1}^n \sum_{j=1}^n u_{ig}^m u_{jg}^m d(\mathbf{x}_i, \mathbf{x}_j)}{2 \sum_{t=1}^n u_{tg}^m} \quad \text{s.t.} \quad u_{ig} \in [0, 1], \sum_{g=1}^k u_{ig} = 1 \quad (2)$$

where u_{ig} represents the generic element of the membership degree matrix $\mathbf{U}$ of order $(n \times k)$, taking values in $[0, 1]$ and representing the membership degree of observation i to cluster g . The parameter m tunes the fuzziness of the obtained solution. As distance measure $d(\mathbf{x}_i, \mathbf{x}_j)$, a Gower's distance [6] is selected given the need to accommodate mixed-type variables. This metric allows the presence of missing values and the assignment of a specific weighting scheme when averaging distances. Its computation results from the following equation:

$$d(\mathbf{x}_i, \mathbf{x}_j)_g = \frac{\sum_{t=1}^r \delta_{ijt} d_{ijt} w_t}{\sum_{t=1}^r \delta_{ijt} w_t}$$

where w_t is the weight assigned to each variable for $t = 1, \dots, r$, $d_{ijt} = 1 - s_{ijt}$ is the distance (or the complement of the similarity s_{ijt}) calculated on the t -th variable, according to its nature (continuous, discrete, categorical, etc.). The case in which equal weights are assigned, i.e., $w_t = 1 \forall t = 1, \dots, r$, results in a balanced contribution of the variables to the overall distance, assuming values in $[0, 1]$ given that it is computed as the average of the standardized distances calculated variable by variable. Finally, the second step to cluster accommodation facilities does not assume an ex-ante-defined number of clusters but selects the best one according to the Fuzzy Silhouette.

3 Data and Results

Situated in an alpine Italian region, the Lower Aosta Valley, composed by 17 municipalities¹, is known for its breathtaking landscapes, outdoor activities, and rich cultural heritage. These aspects make it a destination for niche tourists and a test bed for our methodological proposal. The Lower Aosta Valley's accommodations are often strategically positioned to allow visitors easy access to local attractions such as hiking trails, ski resorts, cultural sites, and historic towns.

The analysis rests on the Booking.com review scores of 44 accommodations in the Lower Aosta Valley, among which are 12 hotels. Even if the average scores (Table 1) are generally high in almost all considered categories (total, service, cleaning, comfort, price-quality ratio, location, and free Wi-Fi), differences in the tourist experience also among the type of accommodations (hotel vs. non-hotel) emerge. Interestingly, the overall score is not an exact linear combination of the ratings assigned to the various satisfaction areas, even if there is a high correlation between them. The total score positively correlates with the score assigned to each service; only free Wi-Fi shows a correlation close to zero (0.027), while the location and the value for money correlate to less than 0.9 with the overall score (0.60 and 0.89). Furthermore, only free Wi-Fi has a negative and close to zero correlation with other services (staff, cleaning, and services). The explored satisfaction areas, such as the quality-price ratio, appear to be of high interest given that they can be traced back to concepts known in the literature, such as price fairness [3]. Therefore, every area is often a proxy of a dimension that internally expresses the tourist's satisfaction on a more complex (economic and psychological) dimension. The latter further justifies our methodological proposal, which addresses the need to identify a comprehensive indicator encompassing various aspects of niche tourists' satisfaction while facilitating quick and straightforward comparisons between accommodations.

The two steps are implemented using the R environment: an own written code for the CPD index, the "fclust" package for the Fuzzy clustering, and the daisy function in the "cluster" package for the Gower's distance are employed.

¹ Arnad, Bard, Challand-Saint-Anselme, Challand-Saint-Victor, Champdepraz, Champorcher, Donnas, Emarèse, Fontainemore, Hone, Issogne, Lillianes, Montjovet, Perloz, Pont-Saint-Martin, Pontboset, Verrès.

Figure 1 shows the CPD index estimates for each accommodation, where the base one (with CPD equal to 1) is the expression of the average service scores. In the second step, the CPD index with categorical variables associated with the level of sustainability (L0, L1, L2, and L3), digitalization (yes/no), municipality, and type of accommodation (hotel or non-hotel) are employed. In the NEFRC, the fuzziness parameter m is set equal to 1.2 (to avoid extremely fuzzy partitions with membership degrees not far from 0.5), and the number of starts equal to 50 to avoid local optima. The optimal number of clusters (from 2 up to 10) is automatically selected using the Fuzzy Silhouette index, which suggests the existence of two clusters. Table 2 and Fig. 1 jointly show the analysis results. In the first cluster, all the accommodation facilities have expressed a non-zero level of sustainability; on average, the value of the CPD index and the total score are lower in this group than in the second. All the facilities in the municipalities of Arnad, Montjovet, Pontboset, Verrès are in the second group, mainly composed of non-hotel accommodation facilities. 46% of the first group is a digitalized hotel. Although further investigation is required, the results are encouraging.

Table 1. Descriptive statistics for the 44 accommodation facilities

		Total	Services	Cleaning	Comfort	Price-quality	Location	Free Wi-Fi
12	mean	8.35	8.31	8.83	8.62	8.51	8.65	6.7
hotel	std	1.07	1.17	0.93	1.12	0.95	0.56	3.33
32	mean	9.17	9.27	9.58	9.44	9.24	8.97	5.15
non-hotel	std	0.43	0.52	0.39	0.42	0.41	0.42	4.36

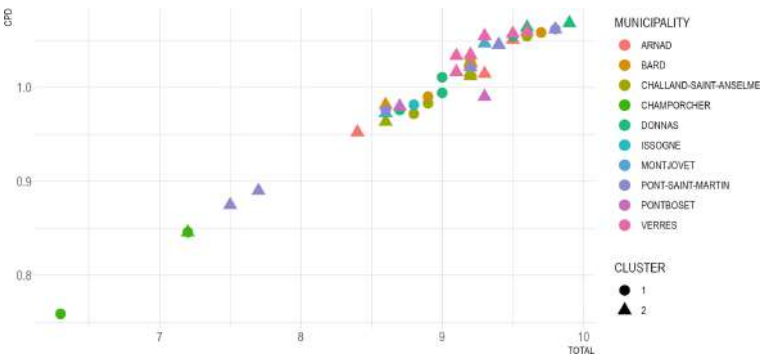


Fig. 1. Estimate of the CDP index vs the total score. Point shapes represent the Fuzzy partition, while the color refers to the municipality.

Table 2. Fuzzy clustering results

	Cluster size	Min memb degree	Max memb. degree	Average memb. degree	N. of unclear. assignment
Cluster 1	15	0.75	1	0.92	0
Cluster 2	29	0.54	1	0.93	0

4 Conclusion

The analysis supports the accommodation facility's management and gives a quick overview of competitors' positioning through clusters and satisfaction index. The latter, capturing deeper aspects of the tourist experience and segmenting the supply, allows for the summary of information that cannot otherwise be directly included within econometric models, given the small sample size characterizing niche tourism destinations. Further research could: i) carry out a municipality aggregation to grasp territorial differences; ii) exploit this approach within a multi-period perspective to register variation over time in nice tourist satisfactions; iii) evaluate alternative weighting schemes in the Gower's distance.

Notes and Comments. This paper is part of the NODES project, which received funding from the MUR-M4C2 1.5 of PNRR and was funded by the European Union-NextGenerationEU (Grant agreement no. ECS00000036). We thank the Aosta Chamber of Commerce for providing us with some data.

References

1. Puh, K., Bagic Babac, M.: Predicting sentiment and rating of tourist reviews using machine learning. *J. Hospital. Tourism Insights* **6**(3), 1188–1204 (2023)
2. Novelli, M., Cheer, J.M., Dolezal, C., Jones, A., Milano, C. (eds.): *Handbook of Niche Tourism*. Edward Elgar Publishing (2022)
3. Alderighi, M., Nava, C.R., Calabrese, M., Christille, J.M., Salvemini, C.B.: Consumer perception of price fairness and dynamic pricing: evidence from Booking.com. *J. Bus. Res.* **145**, 769–783 (2022)
4. Rao, D.P., Hajargasht, G.: A stochastic approach to the computation of purchasing power parities in the International Comparison Program (ICP). *J. Economet.* **191**(2), 414–425 (2016)
5. Runkler, T.A.: Relational fuzzy clustering. *Advances in Fuzzy Clustering and Its Applications*, pp. 31–52 (2007)
6. Gower, J.C.: A general coefficient of similarity and some of its properties. *Biometrics* 857–871 (1971)
7. Nava, C., Osti, L., Zoia, M.G.: From quantity to quality: capturing higher spending markets through a segmentation of travelers' expenditure. *J. Travel Res.* **62**(1), 75–90 (2023)



The Efficacy of Social Distancing Measures in Controlling the Spread of SARS-CoV-2: A Retrospective Study in Italy

Giorgia Bazzurri¹, Maura Mezzetti¹, and Ilia Negri²(✉)

¹ Department of Economics and Finance, Università degli Studi di Roma Tor Vergata, Via Columbia 2, 00133 Rome, Italy

maura.mezzetti@uniroma2.it

² Department of Economics, Statistics and Finance, Università della Calabria, Via P. Bucci, 87036 Rende, CS, Italy

ilia.negri@unical.it

Abstract. Using a Bayesian convolution model, we investigate the impact of government policies regarding social distancing measures in Italy on the transmission of the COVID-19 epidemic. Our methodology entails reconstructing case numbers using observed death rates in all Italian regions. Additionally, we analyze mobility data in conjunction with government-imposed closures. Our objective is to discern the efficacy of these interventions in curtailing the spread of the virus, while also delving into the incurred costs of implementing such measures.

Keywords: Bayesian convolution model · Epidemiological surveillance · COVID-19 · Social distancing

1 Introduction

The paper aims to investigate the impact of various non-pharmaceutical interventions (NPIs) on the spread of COVID-19 and mortality rates in Italy, using a modified statistical model introduced by [1]. Italy, one of the first European countries affected by sustained local transmission of COVID-19, witnessed significant mortality, with 28,238 reported deaths by May 1st, 2020. To mitigate the epidemic, the Italian government implemented several NPIs, including school closures, social distancing measures, and a nationwide lockdown, aiming to prevent healthcare systems from being overwhelmed and to allow time for medical advancements.

As the epidemic worsened in November 2020, the Italian government, under Giuseppe Conte's leadership, adopted a regionally differentiated system of restrictions to mitigate the spread of the virus, aiming to avoid a nationwide lockdown similar to the one imposed in March and April 2020. The regulations underwent changes over several weeks, with each region's color designation based

on weekly case incidence and ICU/hospital occupancy rates. Four colors indicated varying rules: white (minimal restrictions, only masks required), yellow, orange, and red. In yellow zones, movement restrictions were enforced from 10 pm to 5 am, with elementary schools open unless cases emerged. Secondary schools operated at 50% capacity, and certain activities like gyms and theaters were suspended. Dining establishments could serve until 6 pm, limited to four people per table. In orange zones, movement outside one's municipality was banned, and cultural institutions were closed. Dining services were suspended, similar to yellow zones, and school regulations varied by region. Red zones imposed strict movement restrictions and suspended retail activities, with schools also closed.

This study utilizes a Bayesian statistical model, originally proposed in [1] and further developed in [2,3], to elucidate disease transmission dynamics. The rest of the paper is organized as follows: in Sect. 2, data used in the application are presented. Section 3 outlines the model's approach, with a particular emphasis on the role of mobility restrictions. The results are then discussed in Sect. 4, and finally, conclusions are presented in Sect. 5.

2 Data

For the data, COVID-19 regional death counts and cases were obtained from Italy's Civil Protection Department, spanning the period from February 24, 2020, to November 19, 2022, as reported in [4].

Additionally, the publication of mobility data by Google [5] offers a valuable tool for assessing the effectiveness of measures implemented in response to COVID-19 spread. These data provide detailed insights into population-wide changes in movement patterns, serving as a proxy for shifts in behavior and enabling the measurement of SARS-CoV-2 transmissibility. Our analysis utilizes Google's location-based data, in particular summarized in six many categories: (i) retail and recreation, (ii) groceries and pharmacies, (iii) parks, (iv) transit stations, (v) workplaces, and (vi) residential. Google Mobility data serve as a significant proxy measure for social distancing. The Google data are available for each region and are quantified relative to the maximum mobility measured prior to the pandemic WHO declaration. Data are available from February 15th 2020 to the end of 2021. We reconstruct the restrictions on all Italian regions from November 11, 2020, to June 8, 2021, based on information from [6].

3 Model

Let $D_{t,m}$ represent the observed number of deaths for days $t \in \{1, \dots, n\}$ and region $m \in \{1, \dots, M\}$. Assume that $D_{t,m}$ is distributed according to a Negative Binomial distribution:

$$D_{t,m} \sim \text{NegativeBinomial}(d_{t,m}, \phi). \quad (1)$$

The notation in (1) assumes $E(D_{t,m}) = d_{t,m}$, and $Var(D_{t,m}) = d_{t,m} + \frac{d_{t,m}^2}{\phi}$. The parameter ϕ is a parameter modelling the variability in the distribution and a prior positive Gaussian distribution is assumed

$$\phi \sim N^+(0, 5). \quad (2)$$

The expected number of deaths $d_{t,m}$ within a specific locality m at time t can be modeled as a convolution function, as described in Equation (3). This function involves prior days' infection counts, the time from illness onset in a primary case to illness onset in a secondary case, and the dynamically changing reproduction number R_t . This crucial epidemiological metric signifies the average number of new infections stemming from each infected case at time t throughout their period of contagion.

The convolution quantity is then multiplied by the infection fatality rate (*ifr*), that is the overall probability of dying given infection. In our model, we will allow the ifr_m to be region specific and to have some additional noise around this. The noise is multiplicative and distributed as $N(1, 0.1)$. This noise distribution and the prior distribution given by (2) are chosen according to [1]. The quantities ifr_m are download from [7] as provided by Supplementary Material of [1], and are deduced from previous work, [8, 9].

The distribution of the time elapsed between infection and death, π , is assumed to be the sum of two independent random variables: the incubation period (from infection to onset of symptoms or infection-to-onset) and the time between onset of symptoms and death (onset-to-death). The infection-to-onset random variable is Gamma distributed with a shape parameter of 1.35 and a scale parameter of 3.77. Meanwhile, the onset-to-death random variable is also Gamma distributed with a shape parameter of 4.94 and a scale of 3.6. Therefore, the infection-to-death distribution is given by the sum of the two Gamma distributions.

Finally, the expected number of deaths $d_{t,m}$, on a given day t , for region, m , is given by the following discrete convolution sum, that is able to reconstruct the number of cases $c_{t,m}$:

$$d_{t,m} = ifr_m \sum_{\tau=0}^t c_{\tau,m} \pi_{\tau_{t-\tau},m}. \quad (3)$$

The number of cases allows the estimation of the quantity $R_{t,m}$, the effective reproduction number. The effect of each of the p interventions is assumed to be multiplicative. $R_{t,m}$ is therefore a function of the k -th intervention indicators $X_{k,t,m}$ in place at time t in region m :

$$R_{t,m} = R_{0,m} \exp\left(-\sum_{k=1}^p (\alpha_k + \beta^m) X_{k,t,m}\right). \quad (4)$$

The exponential form in (4) was employed to ensure the positivity of the reproduction number, with $R_{0,m}$ constrained to be positive as it appears outside the

exponential. The impacts α_k are shared between all m regions, and therefore, they are informed by all available data and represents the quantities we are interested in. The β^m terms represent the fixed effect for region m .

4 Results

Different models have been fit, considering both observed deaths and 7-day moving average values. This approach helps mitigate differences in daily reported data across various days of the week. The evaluation included assessing the effects of mobility data and non-pharmaceutical interventions on the spread of COVID-19.

The google mobility data are, as expected, correlated, and various combinations, such as the average of different mobility data, are explored. Regional-specific effects, indicated in (4) with β^m , can be covariate-specific, and denoted as β_k^m . The different models fitted are compared in terms of the sum of residuals and convergence. From the results obtained by comparing the considered models, variables related to restrictions appear to be more significant than variables related to mobility provided by Google in explaining the spread of COVID-19. This could be explained by the fact that the values reported by Google are influenced by many factors and may not accurately represent the situation for the entire population of a region.

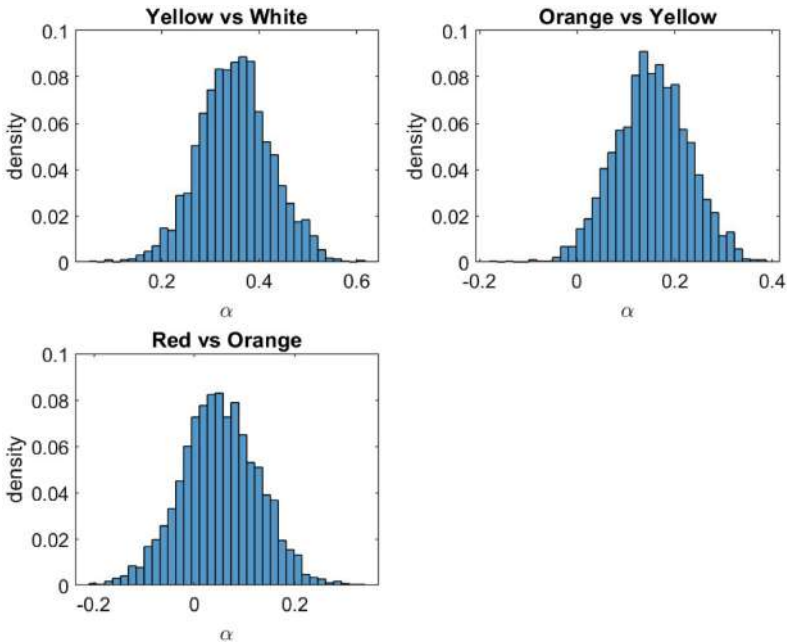


Fig. 1. Posterior distributions for parameter α_k in (4) considering the *colour* decision

In Fig. 1 the posterior distribution of parameter α_k are reported. The main result is the weak effect of the red restrictions respect to the orange one, suggesting that the cost of social distancing measures, such as school closures, may outweigh the benefits in preventing the spread of epidemics. The median values with the 95% credible intervals for the three parameters are 0.3488 (0.2044, 0.4982) for α_1 (yellow vs white), 0.1534 (0.0087, 0.3040) for α_2 (orange vs yellow), and 0.0488 (-0.1019, 0.2007) for α_3 (red vs orange), respectively.

In Fig. 2, the posterior estimates with the 95% credible intervals for the parameters β_k^m , each representing a regional effect, are depicted. This illustrates the diversity in regional responses to the imposed restrictions, highlighting the presence of regional unexplained heterogeneity.

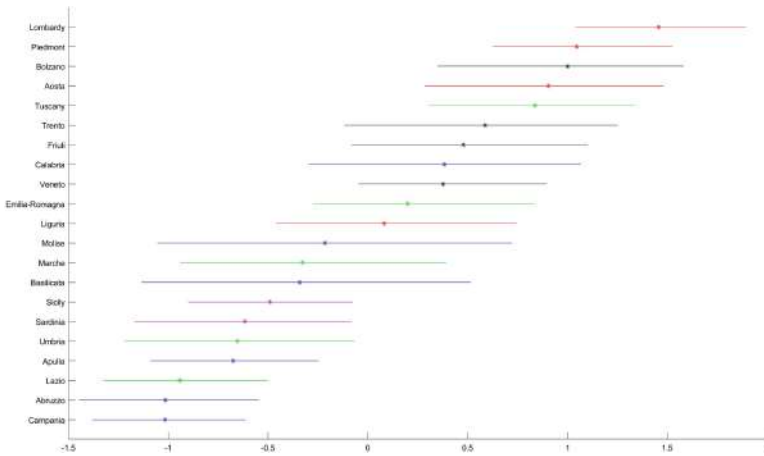


Fig. 2. Posterior estimates for parameter β^m in (4) considering the *colour* decision, value region specific. Black lines indicate regions in North East Italy, red lines indicate regions on North West, green lines for Center, blue lines for South of Italy and finally magenta lines for Islands. (Color figure online)

5 Conclusion

The current study involved modifying a simulation model to assess the impact of various social distancing measures on the spread of COVID-19 in Italy and its regions. The model considered both mobility data and imposed restrictions, with a focus on evaluating the effectiveness of these measures. Our analysis confirms the limited effectiveness of the strongest restrictions, raising questions about their overall impact.

In particular, our study underscores the crucial importance of quantifying the impact of school closures. Schools play a crucial role in the education,

socialization, development, and well-being of children and adolescents. The closure of schools, as indicated by our analysis, may not significantly reduce the spread of the epidemic. However, it could have considerable social and economic implications, leading to disruptions in education and potentially contributing to increased mental health issues.

These findings prompt a thoughtful consideration of the delicate balance associated with implementing stringent measures, especially those affecting educational institutions. While the limited effectiveness in curbing the spread of the virus is acknowledged, the broader social and economic impacts, particularly on the younger generation, should be carefully weighed.

The results of our investigation reveal intriguing patterns. By comparing various models, we find that variables associated with imposed restrictions yield more significance in explaining the spread of COVID-19 than mobility variables provided by Google. This underscores the nuanced nature of pandemic dynamics, where governmental interventions play a crucial role in curbing transmission. Our exploration of the regional effects unveils substantial heterogeneity, indicating that different regions respond differently to imposed restrictions. The posterior estimates of parameters, such as α_k and β^m , provide insights into the week-to-week effects of interventions and regional disparities. Particularly noteworthy is the discernible impact of red restrictions compared to orange ones, suggesting potential trade-offs between the costs of social distancing measures, including school closures, and their effectiveness in curtailing the epidemic.

In conclusion, our retrospective study significantly enhances our comprehension of the dynamics of COVID-19 in Italy. It underlines the need for nuanced, region-specific interventions, showing the substantial regional heterogeneity that existed during the analyzed period. The recognition of diverse responses to interventions among regions emphasizes the critical importance of tailoring strategies to local contexts. These insights, gleaned from a reflective analysis, hold enduring value for policymakers and public health authorities.

References

1. Flaxman, S., Mishra, S., Gandy, A., et al.: Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**(7820), 257–261 (2020)
2. Vollmer, M.A.C., et al.: Report 20: using mobility to estimate the transmission intensity of COVID-19 in Italy: a subnational analysis with future scenarios. *MedRxiv* (2020)
3. Nouvellet, P., Bhatia, S., Cori, A., et al.: Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**(1), 1090 (2021)
4. Dipartimento della Protezione Civile, Presidenza del Consiglio dei Ministri. <https://github.com/pcm-dpc/COVID-19>
5. Aktay, A., et al.: Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.0). *ArXiv abs/2004.0* (2020)
6. Un anno di zone colorate. <https://www.youtrend.it/2021/11/10/un-anno-di-zone-colorate/>

7. Code for modelling estimated deaths and cases for COVID19- Imperial. <https://github.com/ImperialCollegeLondon/covid19model>
8. Altman, G., et al.: A dataset of non-pharmaceutical interventions on SARS-CoV-2 in Europe. *Sci. Data* **9**(1), 145 (2022)
9. Verity, R., et al.: Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet. Infect. Dis.* **20**(6), 669–677 (2020)



A Semi-parametric Spatial Model for Zero Inflated Weibull Distributions with Application to Extreme Rainfall Events

Paolo Onorati^(✉) and Antonio Canale

University of Padova, Padova, Italy
{paolo.onorati, antonio.canale}@unipd.it

Abstract. The mathematical theory underlying classical extreme value methods relies on strong assumptions. Such assumptions are often debated, particularly in the context of rainfall extremes. A possible strategy consists of modeling the ordinary events and making inferences on yearly extremes based on the acquired knowledge of the ordinary events distribution. In this paper, we propose a flexible Bayesian semi-parametric model for the daily rainfall. This method is able to induce spatial dependence and improves previous contributions following similar motivations. Posterior inference is conducted via an efficient Markov Chain Monte Carlo method. Specifically, we show that the proposed algorithm is rejection-free and does not require the calibration of any tuning parameter. We illustrate the method of generating maps of the level return of rainfall extreme in the Veneto and Trentino-Alto Adige regions.

Keywords: Bayesian inference · Gaussian processes · Kolmogorov distribution · elliptical slice sampling

1 Introduction

The Fisher-Tippet-Gnedenko theorem is a notable theoretical result that provides powerful tools for statistical inference for extremes. It states that, if an appropriately standardized maximum of a sequence of *i.i.d.* random variables converges in distribution, then its CDF must belong to the so-called Generalized Extreme Value (GEV) family. From an inferential perspective, this result allows modeling block maxima of *i.i.d.* random variables without caring about their distribution; from theory, it is known that the true distribution belongs asymptotically to the parametric family of the GEV. However, for some applications, the hypotheses of the above theorem are violated.

Modeling yearly maxima rainfall is important for preventing water-related natural disasters like floods or river overflows. Nevertheless; in this context, the assumptions of the Fisher-Tippet-Gnedenko theorem do not hold. For instance; the number of rainy days in a year cannot go to infinity because there are at most 366 days per year.

For this reason; in the context of yearly rainfall extreme, some non-asymptotic methods have been developed [3,4]. These rely on inference from ordinary events. In this paper, we focus on the contributions by Zorzetto et al. [9] and Stolf and Canale [8] that

develop Bayesian procedures that are able to induce spatial dependence through linear functions and hierarchical structures. In this study, we propose an extension of the latter approaches. Specifically, we remove the linear structure using Gaussian processes. From a computational perspective, we propose to use elliptical slice sampling for posterior simulation. In this way, our algorithm is rejection-free and there is no need to calibrate the tuning parameters. We show the use of the proposed method by generating maps of level return in a real data analysis.

2 Statistical Model, Prior Distributions and Computational Method

Let $S = S_1 \times S_2 \cdots \times S_p$ be a p -dimensional spatial domain. We define the following random variables:

$$N_j | \pi(s) \sim \text{Binomial} \left(n = 366, p = (1 + e^{-\pi(s)})^{-1} \right), \quad j = 1, 2, \dots, T,$$

that is, $N_j(s)$ is the number of rainy days for the spatial points s in the j -th year. Thus; for each $j = 1, 2, \dots, T$, we define the following random variables:

$$W_{i,j}(s) | (\gamma(s), \delta(s)) \sim \text{Weibull} \left(\exp(\gamma(s)), \exp(\delta(s)) \right), \quad i = 1, 2, \dots, n_j,$$

say, $W_{i,j}$ is the magnitude of the rainfall for the spatial points s in the j -th year at the i -th rainy day.

We model the parameters of the Binomial and Weibull distributions in a nonparametric way using independent Gaussian processes; that is,

$$\pi(s) \sim \text{GP}(0, k_\pi(s, s')), \quad \gamma(s) \sim \text{GP}(0, k_\gamma(s, s')), \quad \delta(s) \sim \text{GP}(0, k_\delta(s, s')).$$

For all processes, we use the so-called *automatic relevance determination* (ARD) kernel, i.e.

$$k(s, s') = \sigma^2 \exp \left(-\frac{1}{2} \sum_{h=1}^p \frac{(s_h - s'_h)^2}{\lambda_h^2} \right). \quad (1)$$

From (1), it is apparent that we have $p + 1$ hyper-parameters for each process; that is, $(\sigma_\pi^2, \lambda_{\pi,1}^2, \lambda_{\pi,2}^2, \dots, \lambda_{\pi,p}^2)$, $(\sigma_\gamma^2, \lambda_{\gamma,1}^2, \lambda_{\gamma,2}^2, \dots, \lambda_{\gamma,p}^2)$ and $(\sigma_\delta^2, \lambda_{\delta,1}^2, \lambda_{\delta,2}^2, \dots, \lambda_{\delta,p}^2)$. For these, we assume the following prior distributions:

$$\sigma_k^2 \stackrel{i.i.d.}{\sim} \text{Inv.Lomax} \left(\frac{1}{2}, \frac{1}{2} \right), \quad k = \pi, \gamma, \delta,$$

$$\lambda_{k,h}^2 \stackrel{i.i.d.}{\sim} \text{Lomax}(1, 1), \quad h = 1, 2, \dots, p.$$

Here $\text{Inv.Lomax}(a, b)$ denotes a random variable whose reciprocal follows a Lomax distribution with shape a and scale b . Notice that, the prior for σ_k^2 puts an unbounded density close to 0 and it is heavy-tailed; that is, it is a proper distribution that resembles the scale-invariant measure. Notably, all of the priors are heavy-tailed; indeed, the prior mean does not exist for all the above parameters. Furthermore, the use of a $\text{Lomax}(1, 1)$

is consistent with Antoniano-Villalobos et al. [1]; here we use independent Lomax priors instead of the multivariate version due to computational simplicity.

From a computational perspective, the prior for σ_k^2 is related to a scale mixture of inverse gamma distributions; indeed, $\sigma_k^2 | \tau_k^2 \sim \text{Inv.Gamma}(1, \tau_k^2)$ and $\tau_k^2 \sim \text{Gamma}(1/2, 1/2)$. This implies that the full conditionals of σ_k^2 and τ_k^2 are still inverse gamma and gamma random variables respectively. Furthermore, it is worth noting that all the other priors are related to a Gaussian distribution. Indeed, $\pi(s)$, $\gamma(s)$, and $\delta(s)$ are Gaussian; and Lomax priors are inherently linked to a scale mixture of Gaussians. In fact, it is easy to show that the logarithm of a Lomax(1, 1) is distributed as a standard logistic distribution; additionally, a logistic distribution admits a representation as a scale mixture of Gaussians if the mixing density is distributed as suitable transformation of a Kolmogorov distribution [7]. This implies $\log(\lambda_{k,h}^2) | \psi_{k,h}^2 \sim N(0, \psi_{k,h}^2)$, and the density of $\psi_{k,h}^2$ is

$$f_{\psi_{k,h}^2}(x) = \sum_{h=1}^{+\infty} (-1)^{h+1} h^2 \exp\left(\frac{h^2 x}{2}\right).$$

This allows us to exploit the successful elliptical slice sampling [5] to sample from the posterior distribution of the parameters. The elliptical slice sampling is a general method available whenever a density is expressed as a product between a Gaussian density and a generic function; there are no tuning parameters and it is rejection-free. For the $\psi_{k,h}^2$, we use the algorithm proposed in Onorati and Liseo [6].

Therefore, in a Markov Chain Monte Carlo algorithm, we first set the initial values of the MCMC for $\pi(s)$, $\gamma(s)$, and $\delta(s)$ equal to their maximum likelihood estimates, we set $\lambda_{k,h}^2 = 1$ and we sample the remaining variables from their full conditional distributions and then we iterate many times the following updating scheme: (i) update the scaling variables and σ_k^2 , (ii) update $\lambda_{k,h}^2$, and (iii) update the Gaussian processes. Note that, for the first step, it is possible to draw directly from the full conditional distributions while, for the other two steps, we use the elliptical slice sampling. In this way, the simulation scheme just described is essentially a slice sampling within Gibbs.

3 Yearly Maxima Rainfall in Veneto and Trentino-Alto Adige

We consider the rain gauge dataset already analyzed in Dallan et al. [2]. The dataset consists of a 174 heated rain gauge placed in the regions of Veneto and Trentino-Alto Adige; for each weather station, it is collected hourly the quantity of rain for a minimum of 14 years and a maximum of 37 years. The dataset includes also the UTM coordinates and the elevation above sea level. Notice that, this dataset contains a mountainous region known for its complex rain patterns influenced by the unique regional topography.

In this study, we focus our analysis on daily events from January 2005 to December 2008 leading to 1,461 observations for each of the 176 weather stations. Clearly, we do not have positive rainfall for all days and for all weather stations; thus, the actual sample size is 186,647. We standardize the UTM coordinates and then we generate 5,000 posterior draws using the MCMC algorithm described in the previous section for computing return levels associated with spatial points covering the entire grid of interest; that is the Veneto and Trentino-Alto Adige regions. Therefore, given a posterior

sample, we can compute the parameters for Weibull and Binomial random variables. Thus, we obtain one replication from all Weibull distributions and then we compute the yearly maximum. We iterate the entire procedure at the end of each complete scan of the MCMC algorithm for obtaining a sample from the posterior distribution of yearly maxima. Figure 1 reports the maps of the 5, 10, 15, 25, 50, and 100 year return levels.

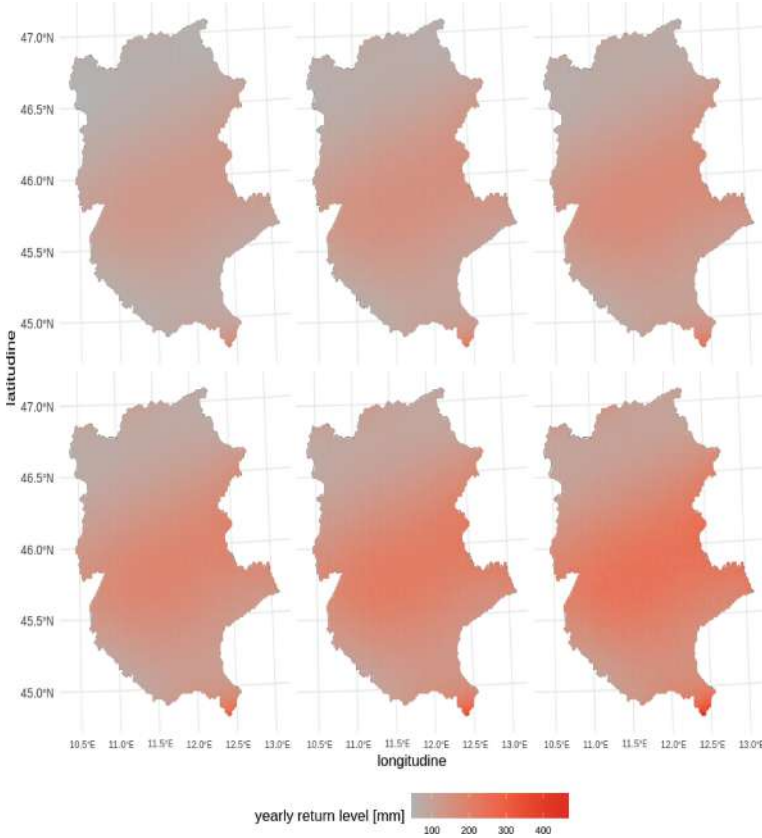


Fig. 1. Maps for the 5 (top-left), 10 (top-center), 15 (top-right), 25 (bottom-left), 50 (bottom-center), and 100 (bottom-left) yearly return levels

4 Conclusion

In this study, we proposed a nonparametric model for yearly maxima rainfall in a non-asymptotic context. Concerning other similar methods, we exploited the use of Gaussian processes in order to overcome the assumption of linearity. We showed that it is possible to construct a simulation algorithm that is easy to use since there are no tuning parameters and it is rejection-free. We illustrated the use of the model by constructing maps of level return for yearly maximum rainfall in real data analysis.

References

1. Antoniano-Villalobos, I., Villa, C., Walker, S.G.: A multidimensional objective prior distribution from a scoring rule. *J. Statist. Plan. Inference* **231**, 106122 (2024)
2. Dallan, E., et al.: How well does a convection-permitting climate model represent the reverse orographic effect of extreme hourly precipitation? *EGUsphere* **2022**, 1–27 (2022)
3. Marani, M., Ignaccolo, M.: A metastatistical approach to rainfall extremes. *Adv. Water Resour.* **79**, 121–126 (2015)
4. Marra, F., Zocatelli, D., Armon, M., Morin, E.: A simplified mev formulation to model extremes emerging from multiple nonstationary underlying processes. *Adv. Water Resour.* **127**, 280–290 (2019)
5. Murray, I., Adams, R.P., MacKay, D.J.C.: Elliptical slice sampling. In: *International Conference on Artificial Intelligence and Statistics* (2009)
6. Onorati, P., Liseo, B.: An extension of the unified skew-normal family of distributions and application to Bayesian binary regression. *arXiv preprint [arXiv:2209.03474](https://arxiv.org/abs/2209.03474)* (2022)
7. Stefanski, L.A.: A normal scale mixture representation of the logistic distribution. *Statist. Probab. Lett.* **11**(1), 69–70 (1991)
8. Stolf, F., Canale, A.: A hierarchical Bayesian non-asymptotic extreme value model for spatial data. *Environmetrics* **34**(7) (2023)
9. Zorzetto, E., Canale, A., Marani, M.: A Bayesian non-asymptotic extreme value model for daily rainfall data. *J. Hydrol.* **628**, 130378 (2024)



Has Immigration Contributed to Population Growth in Spain? A Regional Comparative Analysis

G. Orfao¹ , J. García-Gómez² , A. Del Rey¹ , and E. Parrado³

¹ Department of Sociology and Communication, University of Salamanca, Salamanca, Spain
orfao10@usal.es

² Centre d'Estudis Demogràfics (CED), Barcelona, Spain

³ Department of Sociology, University of Pennsylvania, Philadelphia, PA, USA

Abstract. This paper presents a comparative analysis of the role played by migration in recent population growth trends in six regions of Spain. In recent years, total population has increased in Spain while there has been a strong arrival of immigrants except in the period after the Great Recession. This work contributes to a better understanding of how migration have influenced asymmetrically population growth in six regions of Spain. The databases used are the Residential Variation Statistics and the Natural Movement of the Population for the period 2008–2021. The compensating equation decomposition analysis shows different patterns of population growth across regions: while Spanish regions that show population growth are characterised by having a great positive migration balance of foreign-born individuals, and vice versa, the contribution of foreign-born individuals to births has stopped the fall in the population of those born in Spain. Consequently, all regions show a decline in the group of people born in Spain from native-born parents, except for the Community of Madrid due to the high number of entries of native-born individuals from other regions. Finally, we found differences in the immigrants' contribution to total births according to country of birth.

Keywords: demographic dynamics · migration · regions · Spain

1 Introduction

The aim of this research is to analyse the influence of migration on the population growth of Spain in the last two decades through a comparative analysis at the regional level. Since the beginning of the 21st century, Spain has experienced a strong population growth especially marked by the large arrival of immigrants or “migration boom” during the late 90s and early 2000s (Mujica & Guerra, 2009). Although international immigration was reduced after the economic crisis of 2008, since 2015 the population of those born abroad has increased gradually and unevenly in the different regions of Spain, the Autonomous Communities (CCAA). At the same time, Spain has experienced a strong population aging and a reduction in birth rates, reaching one of the lowest fertility levels worldwide

(Esteve et al., 2021). These demographic challenges highlight the relevance of studying the unequal distribution of immigrants between Spanish regions and its influence on demographic dynamics.

In the debate about the role played by migration in the generational replacement process, some authors emphasise that migration has become a key factor for population growth and replacement in a large part of European countries (Ediev et al., 2014; Ediev & Yüceşahin, 2016; Preston & Wang, 2007). Likewise, Del Rey & Cebrián-Villar (2010) highlight the migration component as the main determinant of this process in Spain, comparing two CCAA such as Madrid and Castilla y León. On the contrary, Serrano-Martínez & García-Marín (2018) stress the ineffectiveness of migration to generate significant changes in the natural demographic trends and characteristics of the Spanish population. By contrary, regional empirical evidence shows a great variety and territorial disparity in relation to demographic processes and behaviours (Carioli et al., 2021; Del Rey & Ortega, 2010). However, as far as we know, comparative analyses at the regional level on the contribution of migration to population growth are practically non-existent.

2 Main Hypotheses

We assume that immigration plays a key role when explaining the recent population growth of some CCAA. Considering the unequal distribution of immigrants according to their origin, we have formulated the following hypotheses:

Hypothesis 1: Migration has a positive effect on the population growth of all CCAA. However, differences are expected across regions. This different contribution is expected in both migration balance and births.

Hypothesis 2: Differences in the population growth and births are expected according to the immigrant population's composition, with a greater contribution of immigration in those CCAA with higher shares of Africans, a group with higher fertility rates.

3 Data and Methodology

3.1 Database

The data used in this research comes from the Residential Variation Statistics (EVR) of the Municipal Register and the Natural Movement of the Population (MNP), provided by the Spanish National Institute of Statistics (INE). These databases allow us to obtain homogeneous and precise annual data at the regional level on births, entries and exits of people from the territory between 2002 and 2021. Another advantage of these databases is the possibility of differentiating both internal and external movements, while, at the same time, we can identify the individuals' country of origin. In this case, we analyse 6 CCAA: the Community of Madrid, Catalonia and Murcia that have experienced population growth, and Castilla y León, Extremadura and Galicia that suffered a decrease during the aforementioned period. Regarding the country of origin, it has been grouped according to different geographical areas, such as Europe, Africa, South America, Central America and the Caribbean, North America, Asia and Oceania.

3.2 Methodology Approach

The methodology approach used in this paper includes a decomposition by country of birth of the commonly known as “compensating equation”. With P_t . Being the total population of a territory in the current period t , the compensating equation allows us to identify the relative contribution of the different factors that influence population growth during a given period, in such a way that:

$$P_t = P_t^\alpha + P_t^\beta = (P_{t-1}^\alpha + P_{t-1}^\beta) + (B_{t-1}^\alpha + B_{t-1}^\beta) - (D_{t-1}^\alpha + D_{t-1}^\beta) + (I_{t-1}^\alpha + I_{t-1}^\beta) - (E_{t-1}^\alpha + E_{t-1}^\beta).$$

being B the total number of births, D the deaths, I the immigration and E the emigration. The contribution of each component to population growth has been decomposed into two factors, one for the native population (α) and another for the immigrant population (β). Furthermore, the sum of births $B_{t-1}^\alpha + B_{t-1}^\beta$ can be decomposed into the births of children whose parents are: (i) both native-born, (ii) both foreign-born, and (iii) one native- and one foreign-born.

4 Rults: Main Contributions, Results and Conclusions

Based on the analysis, certain differences on population growth and demographic trends are observed across different patterns of migration. We present the following conclusions regarding the role of immigration in population growth according to different factors by country of birth in Spain:

1. Although the total population of Spain increased from 2002 to 2021, we found that some CCAA have experienced a decrease in population, such as Castilla y León, Extremadura and Galicia (see Table 1 in Appendix). On the contrary, the population has increased in Catalonia, the Community of Madrid and Murcia. These differences in population growth have been affected, among other factors, by the arrival of foreign-born population. Although we observe an increase in the immigrant population in all CCAA, this rise is concentrated in those regions with a greater population growth. In this sense, immigration has contributed to the growth of the population in Spain while it helped slowing the decline of population in Castilla y León, Extremadura and Galicia.
2. In relation to migration balances, we observe a similar evolution in the six CCAA analysed, with a steady or negative balance after the Great Recession and a subsequent increase starting between 2013 and 2015. When comparing the external balances, we see how the attraction of foreign-born is especially important in the regions where the population has grown (see Table 1 in Appendix). However, while in the CCAA where the population has decreased, we observe a continued internal departure of individuals born in Spain, especially young people, in those where the population has increased, we find two different patterns. In Madrid and Murcia there has been a relative strong arrival of individuals born in Spain since 2008 until the COVID-19 pandemic, while in Catalonia the arrival of national-born individuals is scarce.

3. The decomposition of the compensating equation based on the individuals and parents' country of birth shows a growing trend in the population born in Spain from foreign-born parents, which show an increasing relative weight in all CCAA (see Table 1). Moreover, we observe how in all regions the population of people born in Spain from native-born parents has decreased continuously since 2008, especially in Catalonia. It is concluded that immigration has contributed to population growth due to both the increase in the foreign-born population and its contribution to the number of births.
4. Those CCAA, such as Catalonia and Murcia, where the percentage of immigrants born in Africa and Central America is higher, show a greater growth in the share of individuals born in Spain from foreign-born parents (see Table 1 in Appendix). When weighting the relative change of the number of births per individual by the origin population, this group also reveals an increasing trend in the contribution to births of individuals from these countries (See Fig. 1 in Appendix). Therefore, the composition of the immigrant population also influences the natural balance of the recipient population.

These conclusions support most of the previous evidence on the effects of immigration in relation to the demographic processes of the receiving countries or regions. In a context where the Spanish population is characterized by continuous aging and reduced birth rates which have led to a negative natural balance, our results show the role that the immigrant population has in favouring generational replacement and population growth. Furthermore, considering the movement of young people towards large urban centers and towards certain CCAA, they show the relevance that attracting certain foreign population can have in regions with very low fertility rates and population decline.

Appendix: Tables and figures.

Table 1. Population growth and relative contribution to total change of each component measured as percentage points by regions (CCAA), 2008–2021.

Geographic Area	Period	Total population change	Total natural change	Total migration balance	Relative contribution to total change of each component measured as percentage points*									
					Native-born					Foreign born				
					Vital events			Net migration		Vital events			Net migration	
					Natural change	Births	Deaths	Total	External	Internal	Natural change	Births	Deaths	Total
Community of Madrid	2008–2013	126,838	164,630	–37,792	35.45	70.42	97.97	–53.54	–80.18	246.00	64.55	29.58	2.03	–46.46
	2014–2019	264,566	83,161	181,405	2.12	71.63	96.57	18.32	–35.91	89.81	97.88	28.37	3.43	81.68
	2020–2021	9,462	–13,791	23,253	–210.08	67.63	97.10	–87.77	–11.07	–81.82	110.08	32.37	2.90	187.77
Catalonia	2008–2013	28,278	114,283	–86,005	–7.96	68.93	97.84	–54.58	–46.44	–95.66	107.96	31.07	2.16	–45.42
	2014–2019	151,958	19,842	132,116	–424.01	67.39	97.65	–28.18	–34.02	38.21	524.01	32.61	2.35	128.18
	2020–2021	19,168	–21,286	40,454	–195.53	61.16	97.01	–31.40	–11.36	–112.65	95.53	38.84	2.99	131.40
Murcia	2008–2013	29,967	38,674	–8,707	41.98	73.15	96.65	–19.26	–41.72	135.18	58.02	26.85	3.35	–80.74
	2014–2019	26,021	22,304	3,717	15.60	73.14	96.32	–126.88	–191.14	89.54	84.40	26.86	3.68	226.88
	2020–2021	8,315	1,400	6,915	–157.64	69.11	94.73	–10.02	–13.92	68.71	257.64	30.89	5.27	110.02
Castilla y León	2008–2013	–41,473	–35,810	–5,663	–145.02	83.27	99.40	–345.88	–23.34	–75.29	45.02	16.73	0.60	245.88
	2014–2019	–87,434	–61,451	–25,983	–119.81	84.05	99.33	–125.10	–106.48	–78.00	19.81	15.95	0.67	25.10
	2020–2021	–14,567	–22,570	8,003	–110.36	80.77	99.22	19.03	–13.81	84.29	10.36	19.23	0.78	80.97
Galicia	2008–2013	–9,978	–40,194	30,216	–134.83	86.37	99.26	9.59	–19.41	61.20	34.83	13.63	0.74	90.41
	2014–2019	–45,639	–64,201	18,562	–117.29	86.80	99.21	–40.02	–18.36	–59.35	17.29	13.20	0.79	140.02
	2020–2021	–5,496	–17,497	12,001	–112.48	83.59	98.98	30.31	–2.74	85.74	12.48	16.41	1.02	69.69
Extremadura	2008–2013	2,176	–3,683	5,859	–212.06	91.31	99.47	1.23	–17.08	130.90	112.06	8.69	0.53	98.77
	2014–2019	–30,651	–13,299	–17,352	–125.82	91.36	99.46	–95.51	–246.13	–87.23	25.82	8.64	0.54	–4.49
	2020–2021	–3,941	–5,697	1,756	–111.44	90.15	99.41	–1.94	–17.10	90.69	11.44	9.85	0.59	101.94

* For each component, the relative contribution for both native- and foreign-born is calculated with respect the total contribution of that component for the overall population. This calculated as: $C_{i,t}^\alpha / (C_{i,t}^\alpha + C_{i,t-1}^\beta)$, for each component i *Source:* Authors' elaboration based on EVR and MNP data.

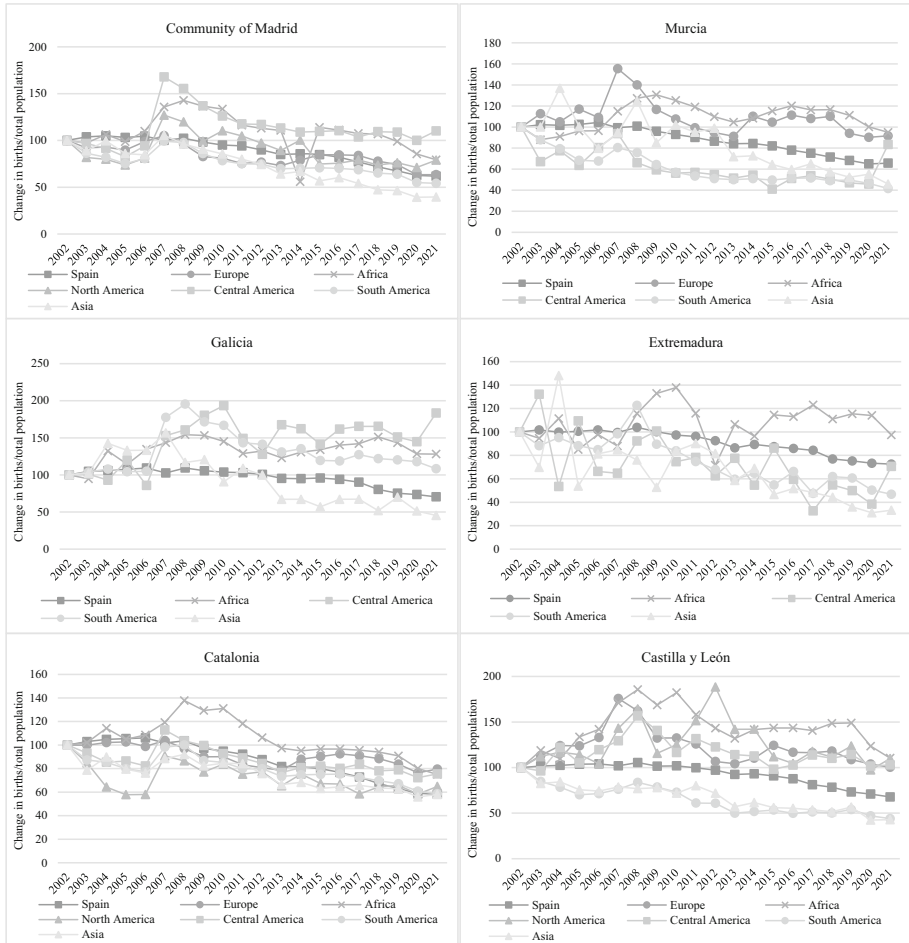


Fig. 1. Relative change in the ratio of births by total population and origin across all regions (CCAA), 2002–2021. Note: Being I the initial year (2002) and P the total population, the relative change in the ratio of births (B) by total population for year t and origin k is calculated as follows: $RC_t^k = (B_t^k/P_t^k)/(B_I^k/P_I^k)$. North America have not been included for Galicia, Extremadura and Murcia, and Europe for the first two due to the small number of individuals from these regions. *Source:* Authors' elaboration based on MNP data.

References

- Carioli, A., Recaño, J., Devolder, D.: The changing geographies of fertility in Spain (1981–2018). *Investigaciones Regionales* **2021**(50), 147–167 (2021)
- Del Rey, A., Cebrián-Villar, M.: Population replacement and migration in two Spanish regions during the twentieth century. *Population* **65**(3), 481–497 (2010)
- Del Rey, A., Ortega, J.A.: The Impact of Migration on Birth Replacement-The Spanish Case. In: *Demographic Aspects of Migration*, pp. 97–121. VS Verlag für Sozialwissenschaften, Wiesbaden (2010)

- Ediev, D., Coleman, D., Scherbov, S.: New measures of population reproduction for an era of high migration. *Popul. Space Place* **20**, 622–645 (2014)
- Ediev, D., Yüceşahin, M.: Contribution of migration to replacement of population in Turkey. *Migr. Lett.* **13**, 377–392 (2016)
- Esteve, A., Lozano, M., Boertien, D., Mogi, R., Cui, Q.: Three Decades of Lowest-Low Fertility in Spain, 1991–2018. Centre d'Estudis Demogràfics, Barcelona. Spain (2021)
- Mujica, J.D., Talavera, R.G.: The demographic transition and immigration flows in Spain: a close relationship. *Geogr. Rev. Japan Series B* **81**(1), 4–15 (2009)
- Preston, S., Wang, H.: Intrinsic growth rates and net reproduction rates in the presence of migration. *Popul. Dev. Rev.* **33**, 357–666 (2007)
- Serrano-Martínez, J.M., García-Marín, R.: Demographic stagnation and decline in Spain: A cause for concern? *Bull. Geogr. Soc.-Econ. Ser.* **39**, 129–145 (2018)



A Fuzzy Clustering Approach to Detect the Number of States in Markov Switching Models

Edoardo Otranto¹ and Luca Scaffidi Domianello²(✉)

¹ Department of Economics, University of Messina, Messina, Italy

² Department of Economics and Business, University of Catania, Catania, Italy
luca.scaffidi@unict.it

Abstract. Markov switching models have received increasing attention in time series analysis due to their ability to detect latent states in the dynamics of the analyzed variables. Generally, the number of states is fixed a priori because it is impossible to apply classical statistical tests, due to the problem of disturbance parameters present only under the alternative hypothesis. In the present contribution we show, via Monte Carlo simulations, that fuzzy clustering can mimic the parametric inference on the regimes derived from the estimated model; furthermore, typical indices used in clustering to determine the number of groups can be used to identify the number of states. The procedure is very simple, considering that it is performed (non-parametrically) independently of the data generation process and that the validation criteria we use are available in most statistical packages. A final application on real data completes the analysis.

Keywords: Nuisance parameters · groups identification · Monte Carlo simulations · Markov chains

1 Introduction

Markov Switching (MS) models [4] have received increasing attention in time series analysis with several applications in economics [7], finance [2], and neuroscience [10] to name a few fields. The main advantage consists in the possibility to take into account several latent states, interpreted as particular regimes (for example, expansion and recession in the business cycle, quiet and turmoil periods in the financial markets), whose dynamics can be described by an ergodic Markov chain. Once we estimate the model, it is possible to make inference on the unobserved state, through the smoothed probability associated with each state. An open issue is the identification of the number of states, which cannot be determined with classical tests for the problem of nuisance parameters present only under the alternative hypothesis. Consequently, in empirical works, the number of states is generally fixed a priori equal to 2 or 3.

In statistical terms the problem can be summarized as follows: when the model under the null hypothesis is an MS model with k states and the model under the alternative is a model with $k + 1$ states, the transition probabilities referring to the state $k + 1$, are not identified under the null hypothesis (see, for example, [8]). As known, in this framework the classical tests, such as Likelihood Ratio, do not follow the standard distributions and the true distribution is unknown.

Specific tests for MS models were developed by [3, 8]. [8] does not derive the exact asymptotic distribution of the statistic, but the upper limit, therefore an approximate distribution; as a consequence, the test proposed is conservative. Furthermore, the distribution is not standard but depends on the model adopted. In particular, to perform the Supremum test, it is necessary to consider several combinations of the parameters constrained under the null hypothesis and the nuisance parameters, with the possibility of applying it only for simple MS models for the high computational costs involved. [3] extends the Supremum test proposed by [9] by excluding the extreme values of the transition probabilities from the range of possible values.

A very practical idea is to identify the number of states before the estimation step, emphasizing the fact that in MS models each observation is generated by a mixture density, with the number of components equal to the number of states of the Markov chain. This approach was developed in a nonparametric Bayesian framework by [11], where the posterior distribution of the number of states is derived using the Gibbs sampler, a well-known Monte Carlo Markov Chain (MCMC) algorithm. This intuition is the basis of our proposal.

First, we note a similarity between the inference on the regime in MS models and the grouping derived from fuzzy clustering [1]. Through this last approach, statistical units can belong to the k clusters with different membership degrees. The intuition is that there is an “assonance” between the probability of a certain state at time t derived from the estimation of MS models and the membership degree of the fuzzy approach. Both approaches can assign a probability to each observation of the analyzed time series and the inference on the state (i.e. the correct assignment of an observation to the state) is more effective the greater this probability is in both approaches. Our first analysis consists of verifying whether the fuzzy approach provides a grouping similar to the one derived from the inference on regimes of the MS models, using the same assignment criterion (the unit is assigned to the group corresponding to the mode). This exercise is performed using Monte Carlo experiments. After verifying the similarity of the results in terms of state inference, we verify whether the typical indices used to select the number of clusters are able to identify the true number of states. We use the fuzzy method as a separate step from the estimation procedure because the aim is to identify the number of states before the estimation step; in practice, the fuzzy procedure is part of the model identification procedure, as in [11], but unlike this, it is very fast and uses tools implemented in the main statistical routines.

The structure of the paper is as follows: in Sect. 2 we present a large set of Monte Carlo experiments to verify the ability of fuzzy clustering to reproduce the inference on the states of MS models and the performance of several indices in detecting the number of states. Section 3 applies the proposed procedure to a real dataset, referring to the Gross Domestic Product (GDP) series of the U.S. Some final remarks will conclude the paper.

2 Monte Carlo Evidence

Let us consider a time series y_t with $t = 1, \dots, T$. The Data Generating Process (DGP) is represented by the following $MS - AR(p)$ model with k states:

$$y_t = \mu_{s_t} + \sum_{i=1}^p \phi_i (y_{t-i} - \mu_{s_{t-i}}) + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad (1)$$

where, $\phi = (\phi_1, \dots, \phi_p)'$ is a vector containing the autoregressive parameters, ε_t ($t = 1, \dots, T$) are independent disturbances with zero mean and constant variance (σ^2). Under stationarity constraints, μ_{s_t} , is a time-varying unconditional mean of the process, depending on an unobservable discrete random variable s_t ,¹ which can assume values $1, 2, \dots, k$, and whose dynamics is driven by an ergodic Markov chain, with elements of the transition probability matrix $\mathbf{P}$:

$$p_{ij} = Pr(s_t = j | s_{t-1} = i), \quad i, j = 1, \dots, k$$

with $\sum_{j=1}^k p_{ij} = 1$ for each $i = 1, \dots, k$. The transition probabilities p_{ij} are estimated with other parameters, $\mu_{s_t} = (\mu_1, \dots, \mu_k)'$, ϕ and σ^2 by Maximum Likelihood Estimator (MLE). The log-likelihood function ($\log L$) can be derived by using the so-called Hamilton filter [4], a sort of Kalman filter with a discrete random variable in the state equation.

We consider 32 DGPs for the Monte Carlo experiments, given by different combinations of the parameters describing an $MS - AR(p)$, that is the model in Eq. 1.²

For each DGP we generate 1000 time series and we perform two types of analyses: first, considering the number of states known, we verify whether fuzzy clustering provides a similar inference on the states obtained from the estimated MS model; then we verify through some clustering validation indices if the clustering procedure is able to detect the right number of states³.

As said, the inference on the state of MS models is performed by assigning the observation at time t to the state with the highest smoothed probability. In our

¹ The variance and the autoregressive parameters could also depend on s_t ; we will not consider this case in our experiments, but the generalization is obvious.

² Details on the DGPs used for the Monte Carlo exercise are available upon request.

³ The adopted clustering validation indices are: Partition Coefficient (PC), Partition Entropy (PE), Modified Partition Coefficient, (MPC), Average Silhouette Width (ASW), Average Silhouette Width Fuzzy (ASWF), and Xie-Beni (XB).

experiment we estimate the corresponding MS model, derive the smoothed probabilities, assign each observation to a state, and finally compare the clustering to the true one using the Rand index [12].

For the sake of space, we do not report the results, which are available upon request. In summary, the classification obtained through fuzzy clustering, which does not take into account the model that generates the data, is very similar to the smoothed inference obtained using the right model (not known in practical cases).

Having empirical evidence of consistent classifications between the typical parametric MS procedure and the fuzzy nonparametric clustering, the next step is to verify if it is possible to use the usual criteria for detecting the number of clusters to identify the number of states in MS models. We then identify, on the same series generated for the previous Monte Carlo experiment, the number of clusters based on the validation indices described above.

Considering DGPs with 2 states, in general, all indices are able to detect the correct number of states with a high percentage of success, excluding the cases where the differences between the switching means are smaller and the variance is larger. PE and PC show a good performance also in these last two cases, with a success rate close to 100%. The other indices are often able to detect the correct number of states, even with a percentage equal to or very close to 100% when the DGP does not contain AR dynamics.

Cases with 3 states present greater uncertainty in detecting the correct number of groups, especially for small differences in the means, but in those cases, the lower the variance the better the performance. The presence of the autoregressive term in DGP seems to worsen the success rate.

In the 3-state case there is no clear preferred index: PC and PE seem less performing than the other four indices, and especially the ASW index shows the highest score with ASWF being second best.

In summary, the simple fuzzy procedure seems to have an overall good performance, especially when the difference among the means is small and the variance is low.

3 An Empirical Application: Identification of the Number of Regimes in U.S. Business Cycle

The seminal paper that sanctioned the success of MS models concerns the analysis of the business cycle [4]. Generally, the number of states is set a priori equal to 2 because the aim is to interpret them as the two phases of the business cycle (expansion and contraction respectively).

We consider the quarterly annualized U.S. GDP growth rate⁴ from 1947:II to 2020:I, thus resulting in 292 observations. By applying a fuzzy clustering procedure, all six validation indices identify the number of states $k = 3$.

Then, we estimate the model in Eq. 1 with 3 states and without the autoregressive parameters (ϕ), and compare it with the model proposed by [5], that

⁴ Data are taken from the Federal Reserve Economic Data database.

is an MS with 2 states. Estimation results, using the Quasi MLE, are shown in Table 1.

We note that, in the case of 2 states, the expansion periods show an average growth rate of 3.85% and -1.67% in the contraction periods. Persistence in the expansion periods is longer than in contraction periods; the average duration of state i is given by $1/(1 - p_{ii})$ (see [6]), so the average duration of the expansion period is 20 quarters and 3.1 quarters for contraction periods. The first state of the MS(3) case shows a very large mean (8.03), while the other 2 states are more in line with the two states of the MS(2) model. Their interpretation is immediate: state 2 is the expansion phase, state 3 the contraction phase and state 1 captures the highest peaks of the series so that it can be interpreted as a boom period. The average duration of the boom is 3.2 quarters, with 11.1 quarters for the expansion phase and 2.9 quarters for the contraction phase. The decrease of the variance parameter in MS(3) compared to MS(2) is clearly due to the presence of more homogeneous growth rates within each state. Finally, AIC and BIC confirm that the MS(3) model shows a superior performance⁵.

Table 1. Parameter estimates with robust standard errors in parentheses for growth rates series of U.S. quarterly GDP (1947:II-2000:I)

μ_1	μ_2	μ_3	p_{11}	p_{12}	p_{21}	p_{22}	p_{32}	p_{33}	σ	AIC	BIC
MS(2)											
3.85	-1.67		0.95			0.68			3.19	5.43	5.49
(0.34)	(1.91)		(0.03)			(0.13)			(0.19)		
MS(3)											
8.03	2.91	-2.40	0.69	0.31	0.03	0.91	0.11	0.66	2.50	5.32	5.45
(0.73)	(0.44)	(4.00)	(0.10)	(0.10)	(0.08)	(0.08)	(0.56)	(0.42)	(0.47)		

Note: The transition probabilities p_{12} and p_{21} in the 2-state cases are obtained as $p_{ij} = 1 - p_{ii}$ ($i, j = 1, 2; i \neq j$). The transition probabilities p_{13} , p_{23} and p_{31} are obtained as $1 - \sum_{j \neq i} p_{ij}$.

4 Final Remarks

Detecting the number of states k is a crucial task in estimating MS models. This problem is often circumvented, by fixing k a priori, due to the impossibility of applying the classical statistical tests for the problem of nuisance parameters present only under the alternative hypothesis. In this work, we underline the similarity between the idea of inference on the state, obtained from MS models, and fuzzy clustering. After verifying the similarity of the grouping with the two procedures through Monte Carlo experiments, on the same simulated series, we

⁵ AIC= $-(2/T)(\log L - P)$; BIC= $-(2/T)(\log L - P \log T)$, with P the number of parameters.

verified the ability of the main clustering validation indices to detect the correct number of states. Our experiments show a good performance, in particular of two indices (ASW and ASWF), also when the dynamics of the time series is affected by autocorrelation. The performance of the indices increases when the differences between the means within states are greater, the variability is lower, and when the length of time series increases. The results seem to support the possibility of adopting clustering validation indices as a tool to identify, before the estimation step, the number of states of MS models in a non-parametric way. The advantage of this procedure is even more evident if we consider that the calculation of these indices is implemented in the main statistical packages.

In future research, it might be interesting to consider other distance measures, such as that of Mahalanobis, or medoid-based approaches as opposed to the centroid-based ones proposed here. It would thus be possible to compare the different fuzzy clustering algorithms to evaluate which one works better.

Furthermore, the procedure proposed above can be extended in a Time Contrastive Learning (TCL) framework [10] to identify the number of latent states assumed known.

Acknowledgements. Edoardo Otranto acknowledges financial support from the Italian PRIN 2022 grant “Methodological and computational issues in large-scale time series models for economics and finance” (20223725WE).

Luca Scaffidi Domianello acknowledges financial support from PNRR MUR project PE0000013-FAIR.

References

1. D’Urso, P.: Fuzzy clustering. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds.) *Handbook of Cluster Analysis*, pp. 545–566. Chapman and Hall (2015)
2. Gallo, G., Otranto, E.: Forecasting realized volatility with changing average volatility levels. *Int. J. Forecast.* **31**, 620–634 (2015)
3. Garcia, R.: Asymptotic null distribution of the likelihood ratio test in Markov-Switching models. *Int. Econ. Rev.* **39**, 763–788 (1998)
4. Hamilton, J.: A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384 (1989)
5. Hamilton, J., Chauvet, M.: Dating business cycle turning points. In: NBER (2005)
6. Hamilton, J.D.: *Time Series Analysis*. Princeton University Press, Princeton (1994)
7. Hamilton, J.D.: Macroeconomic regimes and regime shifts. In: Taylor, J.B., Uhlig, H. (eds.) *Handbook of Macroeconomics*, Chapter 3, pp. 163–201. Elsevier (2016)
8. Hansen, B.: The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *J. Appl. Economet.* **7**, S61–S82 (1992)
9. Hansen, B.E.: Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64**, 413–430 (1996)
10. Hyvärinen, A., Khemakhem, I., Morioka, H.: Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns* **4**, 1–14 (2023)
11. Otranto, E., Gallo, G.M.: A nonparametric Bayesian approach to detect the number of regimes in markov switching models. *Economet. Rev.* **21**, 477–496 (2002)
12. Rand, W.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971)



Trimmed Symmetric Linear Regression

Simona Pacillo^{1(✉)}, Luca Greco², and George Luta³

¹ University of Sannio, Benevento, Italy
simona.pacillo@unisannio.it

² University Giustino Fortunato, Benevento, Italy
l.greco@unifortunato.eu

³ Georgetown University, Washington, DC, USA
George.Luta@georgetown.edu

Abstract. Symmetric regression deals with a reversible functional relationship between (Y, X) regardless of the direction of regression. Orthogonal, Geometric Mean, Pythagorean, Deming regression are valuable examples. However, contamination in the data can have a seriously adversarial effect on estimation. Therefore, a robust methodology for symmetric regression is proposed, that is built on the basics of Trimmed Least Squares: the reversible regression line is fitted on a subset of the data whereas potential outliers are discarded.

Keywords: Concentration step · Outliers · Reweighting · Subsampling

1 Introduction

Symmetric regression allows to study the functional relationship between a set of variables through a regression model specification, without distinguishing between the response Y and the set of covariates X . In other words, symmetric regression deals with a reversible functional relationship regardless of the direction of regression. Let us consider the case of two variables (Y, X) for simplicity. According to Ordinary Least Squares (OLS), it is possible to consider two different models, depending on which variable is taken as the outcome, and then to predict X from Y starting from a point fitted by estimating the regression of Y on X . Let $\hat{y} = \hat{y}(x) = E[Y|X = x]$ and $\hat{x} = \hat{x}(y) = E[X|Y = y]$. From OLS, we have that $\hat{y}(\hat{x}) \neq \hat{x}^{-1}(\hat{y})$, unless perfect correlation exists among the variables, where $x^{-1}(y)$ denotes the regression of X over Y expressed in y -units. Then, the two models lead to different predictions and can not be reduced from one to another. In contrast, in symmetric regression, we want the equality $\hat{y}(\hat{x}) = \hat{x}^{-1}(\hat{y})$ to hold.

The study of a reversible functional relationships is of interest in many fields, such as physics, economics, social sciences, psychology, biostatistics, astronomy and so on, in all those situations where it is meaningful to consider regression in any direction and there is not clear indication of what is to be considered as response and what not, for instance when comparing a new and a reference assays (see [8, 13] for a review).

Several modification of classical linear regression models are available that extend to the case of simultaneous fitting by all directions. The first kind of symmetric regression is orthogonal regression (OR, [10]). A second solution is geometric mean regression, also known as reduced major axis, diagonal or impartial regression, (GMR, see [14] for a review). Its main improvement over OR is that GMR is scale independent. Among others, further solutions include Pythagorean regression (PR [6]) and Deming regression (DR [2, 9, 11]). The reader is pointed to [5] for a review.

As well as OLS, symmetric regression can be badly affected by the occurrence of several outliers corrupting the sample data. Therefore, there is the need to develop procedures that are resistant to contamination in the data and able to provide robust inferences. In this respect, OR M-estimation was introduced in [15], whereas a weighted OR was proposed in [4]. A procedure based on Least Median of Squares (LMS) was suggested in [7] and applied to orthogonal residuals. Here, we develop a general strategy for robust fitting of a symmetric regression model that is based on the idea of trimming underlying Least Trimmed Squares (LTS, see [12] and references therein). The proposed methodology generalizes the former approaches in a couple of directions. First, it is based on LTS rather than on LMS. Then, it can be applied to different symmetric regression models going beyond OR.

This contribution is structured as follows. In Sect. 2 we give a brief background on symmetric regression. The robust fitting procedure based on trimming is presented in Sect. 3. Some examples are illustrated in Sect. 4.

2 Symmetric Regression

Consider a sample (y_i, x_i) , $i = 1, 2, \dots, n$ and the functional linear regression model

$$\begin{cases} y(x_i) = \bar{y} + \beta(x_i - \bar{x}) \\ x(y_i) = \bar{x} + \frac{1}{\beta}(y_i - \bar{y}) \end{cases}$$

where $(\bar{y}, \bar{x})$ denotes the vector of means and $y(x), x(y)$ denote the conditional means in both directions. Parameters estimation is obtained from the minimization of a suitable objective function of the form

$$Q(\beta) = g(\beta) \sum_{i=1}^n d_{y_i}^2(\beta) \quad (1)$$

where $d_y(\beta) = (y_i - y(x_i))$ is the usual *vertical* distance from the regression line. The objective function (1) can be obtained from a suitable summary of both squared vertical d_y^2 and horizontal $d_x^2 = (x_i - x(y_i))^2 = d_y^2/\beta^2$ squared errors (see [6] for a review). In particular, GMR is defined considering the minimization of the sum of the geometric mean of squared vertical and horizontal distances and corresponds to the choice $g(\beta) = |\beta|^{-1}$. In a different fashion, PR is obtained from the minimization of the sum of the arithmetic mean of squared distances

and $g(\beta) = \left(1 + \frac{1}{\beta^2}\right)$. Regarding OR, it is based on the harmonic mean: minimizing the sum of the harmonic mean of the squared vertical and horizontal distances leads to the same result as minimizing the sum of squared orthogonal distances; in this case $g(\beta) = (1 + \beta^2)^{-1}$. DR is derived from a likelihood approach to an Error-in-Variables model [11]

$$\begin{cases} x_i = u_i + \epsilon_i^x \\ y_i = \alpha + \beta u_i + \epsilon_i^y. \end{cases}$$

Under the assumptions of independence among the errors (ϵ^x, ϵ^y) and constant variance $Var(\epsilon_i^x) = \sigma_x^2$, $Var(\epsilon_i^y) = \sigma_y^2$, the objective function (1) is now characterized by $g(\beta) = (\sigma_y^2 + \beta^2 \sigma_x^2)^{-1}$. In the case in which it is assumed that $\sigma_y^2 = \sigma_x^2$, then we recover OR; under the hypothesis of proportional variances $\sigma_x^2 = k\sigma_y^2$, then $g(\beta) = (1 + k\beta^2)^{-1}$. In practice, when k is not known in advance, it is estimated using multiple measurements within each subject.

3 Methodology

The objective is to obtain a fitted symmetric regression line that is not affected by the occurrence of outliers. The ideal solution consists in discarding outliers and run the analysis on the genuine data only. Here, we focus on a methodology that leads to a robust fit and discard outliers simultaneously. Actually, LTS aims at finding the subset of h points leading to the minimum least squares criterion. As well, in a symmetric regression framework, one is interested in the h -subset minimizing

$$Q_h(\hat{\beta}) = g(\hat{\beta}) \sum_{i=1}^h d_{y_{(i)}}^2(\hat{\beta}) \quad (2)$$

where $\hat{\beta}$ is the fitted slope to the selected h points and $d_{y_{(i)}}^2(\hat{\beta})$ are the squared ordered (vertical) residuals. After a suitable initialization strategy, the fitting process iterates between a concentration (C-) and an optimization (O-) step. In the C-step, based on the current estimate of the slope, we compute residuals and sort them in non decreasing order according to their absolute value. In the O-step, one minimizes (2) over the h points with the smallest absolute residuals. The value $Q_h(\hat{\beta})$ at convergence is a local minimum. A common strategy to find a global minimum is to run the procedure from different starting points obtained using subsampling.

Furthermore, in order to improve the trimmed estimate based on the selected h -subset, one can consider a reweighing step as suggested in [1]. The simplest case is based on the use of hard-rejection weights $w_i = w(|d_{y_i}(\hat{\beta})|/\hat{\sigma})$ with $w_i = 1$ if $(|d_{y_i}(\hat{\beta})|/\hat{\sigma}) < c$, where $\hat{\sigma}$ is a suitable robust scale estimate of the residuals and c is a constant usually equal to a high quantile of the normal distribution.

4 Examples

The first example concerns blinded data from a clinical laboratory. They are stored in the data frame `ferritin2` from the R package `deming` [3]. For each of seven periods, patients were assayed for ferritin content using both the old and new batches of reagent. Here, we consider the data from the second period that contains some outliers artificially added. We considered trimmed and reweighted OR, GMR and PR with an initial rate of trimming equal to 0.25 and reweighting run at a significance level 0.01. The methods lead to almost identical results. The robust fits by reweighted GMR is displayed in Fig. 1. The fitted coefficients are $\hat{\alpha} = -1.207$ and $\hat{\beta} = 0.995$. After the reweighting step, the outliers have been correctly detected and discarded.

The second dataset [15] gives simultaneous pairs of measurements of kanamycin levels in blood samples drawn from 20 babies. Case number two is an outlier. Figure 2 shows reweighted PR along with its non robust counterpart, for an initial rate of trimming equal to 0.25 and reweighting run at a significance level 0.01. The classical fit gives $(\hat{\alpha}, \hat{\beta}) = (-0.733, 1.049)$, whereas the robust fit returns $(\hat{\alpha}, \hat{\beta}) = (-3.937, 1.229)$ and correctly identifies the anomalous point.

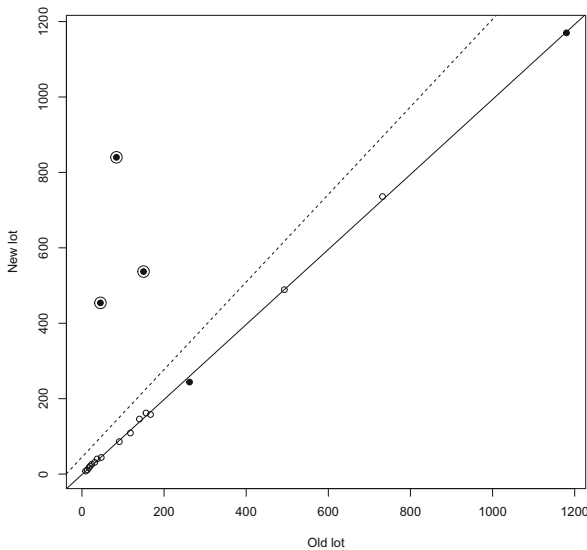


Fig. 1. Fitted symmetric regression line by reweighted (solid line) and classical (dashed line) GMR. Filled black circles denote trimmed points. Points discarded after reweighting are highlighted by a surrounding circle.

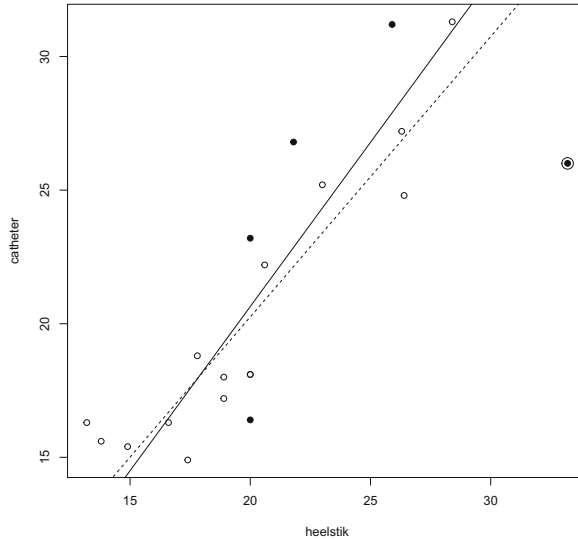


Fig. 2. Fitted symmetric regression line by reweighted (solid line) and classical (dashed line). Filled black circles denote trimmed points. Points discarded after reweighting are highlighted by a surrounding circle.

References

1. Čížek, P.: Reweighted least trimmed squares: an alternative to one-step estimators. *TEST* **22**(3), 514–533 (2013)
2. Deming, W.E.: *Statistical Adjustments of Data*. Wiley, New York (1943)
3. Therneau, T.: Deming: Deming, Theil-Sen, Passing-Bablok and Total Least Squares Regression. R package version 1.4 (2018). <https://CRAN.R-project.org/package=deming>
4. Fekri, M., Ruiz-Gazen, A.: LTS weighted orthogonal regression in the errors-in-variables model. *J. Multivar. Anal.* **88**(1), 89–108 (2004)
5. Francq, B.G., Govaerts, B.B.: Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives. *Chemom. Intell. Lab. Syst.* **134**, 123–139 (1989)
6. Greene, N.: Generalized least squares regression I: efficient derivations. In: *1st International Conference on Computational Science and Engineering* (2013)
7. Hartmann, C., Vankeerberghen, P., Smeyers-Verbeke, J., Massart, D.: LTS orthogonal regression for the outlier detection when comparing two series of measurement results. *Anal. Chim. Acta* **344**(1–2), 17–28 (1997)
8. Isobe, T., Feigelson, E.D., Akritas, M.G., Babu, G.J.: Linear regression in astronomy. *Astrophys. J. Part 1* **364**, 104–113 (1990). ISSN 0004-637X. Research Supported by NASA **364**, 104–113 (1990)
9. Linnet, K.: Estimation of the linear relationship between the measurements of two methods with proportional errors. *Stat. Med.* **9**(12), 1463–1473 (1990)
10. Pearson, K.: On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philos. Magaz. J. Sci.* **2**(11), 559–572 (1901)

11. Ripley, B.D., Thompson, M.: Regression techniques for the detection of analytical bias. *Analyst* **112**(4), 377–383 (1987)
12. Rousseeuw, P.J., Van Driessen, K.: Computing lts regression for large data sets. *Data Min. Knowl. Disc.* **12**, 29–45 (2006)
13. Taagepera, R.: *Making Social Sciences More Scientific: The Need for Predictive Models*. Oxford University Press, USA (2008)
14. Xu, S.: A property of geometric mean regression. *Am. Stat.* **68**(4), 277–281 (2014)
15. Zamar, R.H.: LTS estimation in the errors-in-variables model. *Biometrika* **76**(1), 149–160 (1989)



Analysis of Complex Spatio-Temporal Neuroimaging Signals by Functional Principal Component Analysis

Alessandro Palummo^{1(✉)}, Eleonora Arnone², Letizia Clementi³,
and Laura M. Sangalli¹

¹ Politecnico di Milano, MOX - Department of Mathematics, Milano 20133, Italy
alessandro.palummo@polimi.it

² Università degli Studi di Torino, Department of Management, Turin 10134, Italy

³ Fondazione IRCCS Istituto Neurologico Carlo Besta, SC Neurologia,
Salute Pubblica, Disabilità, Milan, Italy

Abstract. We propose a novel methodology to extract the main sources of variability from a collection of space-time dependent signals observed over complicated geometries. The methodology is developed in the context of functional Principal Component Analysis (fPCA), and proposes an estimation problem combining a rank-one approximation of the data matrix with a roughness penalty. The computed principal components are smooth spatio-temporal functions over the domain of interest, which are easy to interpret and can lead to interesting insights in the spatio-temporal dynamic of the phenomenon under study. The model is applied to the study of neuroimaging data. In particular, we explore the main sources of variability in neuronal connectivity in a population of healthy and pathological subjects, starting from functional Magnetic Resonance Imaging scans.

Keywords: functional principal component analysis · space-time dependent data · neuroimaging data · roughness penalties

1 Introduction

Consider the problem of exploring the main variability patterns in a collection of spatio-temporal signals, acquired over a multidimensional complicated domain. In neuroimaging, an example in this sense is provided by the Blood Oxygenation Level Dependent (BOLD) signal, acquired by functional Magnetic Resonance Imaging (fMRI), which serves as a proxy measure for neuronal activity [13]. The high dimensionality of fMRI data, together with the anatomical complexity of the brain, poses different methodological and computational challenges, which invalidate existing statistical approaches for dimensionality reduction, typically based on some form of multivariate Principal Component Analysis (PCA).

The goal of the present work is to introduce a novel methodology for the study of the main variability patterns inside space-time dependent data. We do so in the

framework of Functional Data Analysis (FDA) [4, 7, 11]. Classical approaches to functional Principal Component Analysis (fPCA) involve a pre-smoothing of the observed data, which are considered as discrete and noisy evaluations of smooth realizations of an underlying random field, followed by an eigen-decomposition of its sample covariance operator. These techniques can lead to unsatisfactory results [1, 8]. Moreover, if pre-smoothing is performed using standard techniques, such as splines expansions, this would ignore the anatomical structure of the brain, leading to artifacts in the estimates.

Our approach, instead, is not based on any form of pre-smoothing and originates from a different literature, first explored in [6] and [8], and is a new addition to the class of Physics-Informed statistical methods, reviewed in [12]. Specifically, the authors in [8] proposed a form of fPCA for the analysis of data located over complicated two-dimensional manifold, such as the cortical surface of the brain. The proposed model employs a penalization term that properly accounts for the geometry of the considered domain in the estimation problem. An extension of this model has been considered in [1, 3] for the analysis of the main variability patterns of the Functional Connectivity maps computed within the gray matter volume. The case of planar domains with spatio-temporal signals affected by large gaps in space and time has been already discussed in [9].

Following the penalized approach explored in [1, 3, 8], we here extend the fPCA methodology for the analysis of space-time dependent data located within volumetric domains with complicated shapes. We do so by considering a proper space-time dependent differential penalty, first introduced in [2] in the context of spatio-temporal smoothing, which is capable to take into account both the anatomical complexity of the brain and the temporal structure of the BOLD signal. The resulting estimated principal components are smooth functions over the spatio-temporal domain of interest. The use of advanced numerical discretization methods, based on finite elements, enables to obtain a computationally tractable formulation of the estimation problem. The resulting technique permits to explore the main modes of variations of space-time dependent neuroimaging signals, while properly complying for the anatomy of the brain.

2 Model

Let $\mathcal{D}$ be a bounded spatial domain in $\mathbb{R}^3$ and $T \in \mathbb{R}^+$ a bounded time interval. Consider a random function $\mathcal{X}$ with values in $L^2(\mathcal{D} \times T)$, the latter being the space of all squared integrable functions on $\mathcal{D} \times T$. Under some regularity conditions on the random function $\mathcal{X}$, we can define the principal components of $\mathcal{X}$ as the eigenfunctions of its covariance operator (see, e.g., [11]). Ideally, we would consider a sample of L realization of the random function $\mathcal{X}$, given by $x_1, x_2, \dots, x_L$ observed continuously and without noise over $\mathcal{D} \times T$, and recover the first principal components of the random function from the spectral decomposition of its covariance operator. In practice, only noisy measurements of these functions are available, at some spatio-temporal locations $\{(\mathbf{p}_i, t_j)\}_{ij} \in \mathcal{D} \times T$. A common approach in FDA is to recover smooth representatives of $x_1, x_2, \dots, x_L$,

by expressing them in terms of some basis expansion, and then recover the sample covariance operator of $\mathcal{X}$ together with its spectral decomposition. In doing so, part of the variability in the original data is lost as a consequence of the smoothing procedure, yielding sub-optimal results.

In the proposed work, we exploit a different characterization of the Principal Components, first explored in [6]. Specifically, for a zero-mean random function $\mathcal{X}$, the first M principal components are, among all M -dimensional orthonormal bases, that providing the best representation of the random function, in the L^2 -sense. Let $f \in L^2(\mathcal{D} \times T)$ be the first Principal Component (PC) function and $\mathbf{s} \in \mathbb{R}^L$ be the corresponding PC scores vector. Denote with $\mathbf{f}_n$ the evaluation of f at the discrete spatio-temporal locations, and let X be the data matrix. The discrete counterpart of the cited characterization implies that the representation of the original data on the first principal component, computed as the rank-one matrix $\mathbf{s}\mathbf{f}_n^\top$, best approximates, in the Frobenius norm sense, the data matrix X . In other words, the pair $(\mathbf{s}, f)$ is such that it minimizes the data loss $\|X - \mathbf{s}\mathbf{f}_n^\top\|_F^2$.

Following the penalized approach discussed in [1, 8], we combine the considered data loss with a proper roughness penalty term, seeking for the smoothness of the estimated PC function f in space and time. As discussed by the authors in [2], we introduce the separable spatio-temporal penalty $\mathcal{P}_{\lambda_D, \lambda_T}(f) = \lambda_D \int_{\mathcal{D}} \int_T (\Delta f)^2 + \lambda_T \int_T \int_{\mathcal{D}} (\frac{\partial^2 f}{\partial t^2})^2$, with Δf denoting the Laplacian of f and $\lambda_D, \lambda_T > 0$ being smoothing parameters. We hence propose to estimate the optimal pair $(\mathbf{s}, f)$ by minimization of the following functional:

$$\begin{aligned} & \arg \min_{f \in L^2(\mathcal{D} \times T), \mathbf{s} \in \mathbb{R}^L} \|X - \mathbf{s}\mathbf{f}_n^\top\|_F^2 + \mathbf{s}^\top \mathbf{s} \mathcal{P}_{\lambda_D, \lambda_T}(f) \\ & \text{Subject to} \quad \|\mathbf{s}\|_2 = 1 \end{aligned} \quad (1)$$

The estimation problem in Equation (1) poses different challenging aspects. First of all, it is an infinite dimensional minimization problem in the unknown f and does not enjoy an analytical solution. Secondly it is non-convex in the unknown pair $(\mathbf{s}, f)$. To tackle these issues, we develop appropriate numerical discretization techniques, coupled with an opportune iterative minimization procedure. Specifically, we express the PC function f with respect to a separable spatio-temporal basis system, obtained by a proper tensorization of a finite element basis [10] built over a triangulation of the spatial domain $\mathcal{D}$ and a cubic B-Spline basis [14] defined over the time interval T ; for details, we refer to [9]. The minimization of the functional in Equation (1) is therefore performed by an iterative two-step procedure, where we alternate the estimation of the PC scores vector $\mathbf{s}$ given f , and the estimation of the PC function f given $\mathbf{s}$, until convergence. The latter proved to be equivalent to a proper smoothing problem for space-time dependent data.

Subsequent PC functions are obtained solving the same problem in Equation (1) on the residualized data matrix $X - \mathbf{s}\mathbf{f}_n^\top$. The smoothing parameters λ_D and λ_T , as well as the optimal number of principal components M , can be selected using the same arguments detailed in [8].

3 Application

To show the validity of the proposed fPCA, we analyze the main variability patterns inside the BOLD signal acquired within the grey matter volume, using resting-state fMRI. Specifically, we employ the preprocessed version of the dataset provided by [5], made available by the Consortium for Neuropsychiatric Phenomics at UCLA, and already analyzed in [3]. As is common in neuroimaging applications, data have been aligned on a common brain atlas, which corresponds to our spatial domain $\mathcal{D}$. In order to apply the discussed fPCA, we consider a tetrahedral mesh discretizing the atlas, on which we can define a proper finite element space. Specifically, the discretized spatial mesh consists of 36K nodes. To each node is then associated a BOLD time-series, made by 34 acquisitions in time, that are filtered to remove high frequency noise. Figure 1 shows the complicated anatomy of the brain volume with a frame of the considered data.

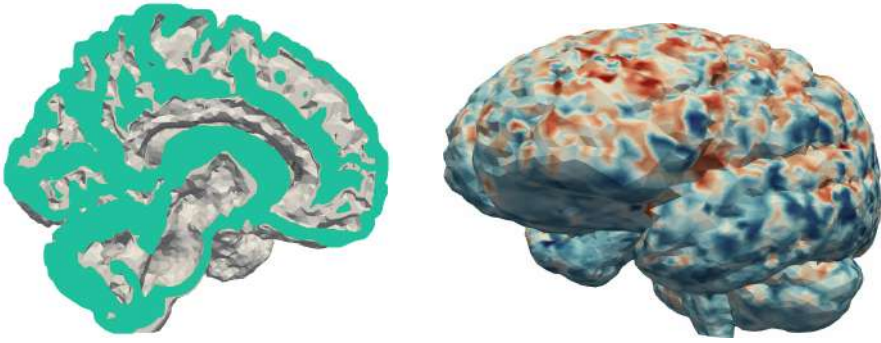


Fig. 1. Left: slice of the brain volume, showing the complicated geometry of the problem under analysis. Right: example of considered data for one patient at a fixed time instant.

The discussed fPCA is able to analyze, at a reasonable computational cost, massive amounts of spatio-temporal dependent data located over domains with non-trivial geometries, characterized by holes and sharp concavities. Differently from other functional approaches, based on the Euclidean metric, the presented methodology properly accounts for the highly non-Euclidean structure of the brain, proved to be of great importance for the correct modeling of the brain functioning. Moreover, the use of a roughness penalty leads to Principal Component functions which are interpretable, making the proposed fPCA a valuable tool for the analysis of complex neuroimaging signals.

Acknowledgements. A. Palummo acknowledges funding by European Union - Next-GenerationEU. We also acknowledge the MUR research project Dipartimento di Eccellenza 2023–2027, Dipartimento di Matematica, Politenico di Milano.

References

1. Arnone, E., Negri, L., Panzica, F., Sangalli, L.: Analyzing Data in Complicated 3D Domains: smoothing, semiparametric regression, and functional principal component analysis. *Biometrics*. **79**(4), 3510–3521 (2023). <https://doi.org/10.1111/biom.13845>
2. Bernardi, M.S., Sangalli, L.M., Mazza, G., et al.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stoch. Environ. Res. Risk Assess.* **31**, 23–38 (2017)
3. Clementi, L., Arnone, E., Santambrogio, M.D., Franceschetti, S., Panzica, F., et al.: Anatomically compliant modes of variations: new tools for brain connectivity. *PLoS ONE* **18**(11), e0292450 (2023). <https://doi.org/10.1371/journal.pone.0292450>
4. Ferraty, F.: Nonparametric functional data analysis. Springer (2006). <https://doi.org/10.1007/0-387-36620-2>
5. Gorgolewski, K., Durnez, J., Poldrack, R.: Preprocessed consortium for neuropsychiatric phenomics dataset. *F1000Research* **6** (2017)
6. Huang, J., Shen, H., Buja, A.: Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.* **2** (2008)
7. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
8. Lila, E., Aston, J., Sangalli, L.: Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.* **10**, 1854–1879 (2016)
9. Palummo, A., Arnone, E., Formaggia, L., Sangalli, L.M.: Functional principal component analysis for incomplete space-time data. *Environ. Ecol. Stat.* (2024). <https://doi.org/10.1007/s10651-024-00598-7>
10. Quarteroni, A.: Numerical Models for Differential Problems. Springer International Publishing, Cham (2017). <https://doi.org/10.1007/978-88-470-5522-3>
11. Ramsay, J., Silverman, B.: Functional data analysis. Springer, New York (2005). <https://doi.org/10.1007/b98888>
12. Sangalli, L.M.: Spatial regression with partial differential equation regularization. *Int. Stat. Rev.* **89**(3), 505–531 (2021)
13. Soares, J.M., et al.: A Hitchhiker’s guide to functional magnetic resonance imaging. *Front. Neurosci.* **10**, 515 (2016). <https://doi.org/10.3389/fnins.2016.00515>
14. Wahba, G.: Spline models for observational data. *Soc. Ind. Appl. Math.* (1990)



Asthma Prediction Tool in Preschool Children: Evidence from Predictive Performance Comparison by Using Innovative Statistical Approach

Alessandra Pandolfo^{1,2(✉)}, Gianluca Sottile², Velia Malizia¹, Omar Shatarat¹, Valentina Lazzara^{1,2}, Vito Muggeo², Giovanni Viegi³, and Stefaina La Grutta²

¹ Institute of Translational Pharmacology, CNR, Palermo, Italy

alessandra.pandolfo@ift.cnr.it

² Department of Economics, Business and Statistics, UNIPA, Palermo, Italy

alessandra.pandolfo@unipa.it

³ Institute of Clinical Physiology, CNR, Pisa, Italy

Abstract. This study aims to develop a non-invasive tool for early asthma diagnosis in preschool children. The tool leverages co-morbidities, environmental factors, and socio-economic determinants. The research compares Traditional Statistical Models, Machine Learning Techniques, and Deep Neural Networks to assess the predictive capabilities of various models against the Leicester tool, which is used as a benchmark. The study's results, evaluated on accuracy and AUC metrics, indicate that the proposed tool outperforms the Leicester model across all applied models. This research highlights the potential of integrating diverse variables for asthma prediction, suggesting directions for future enhancements and underscoring the need for comprehensive evaluations to validate these findings.

Keywords: Asthma · Machine learning · Deep Neural Network · Predictive tool · Preschool Children

1 Introduction

Asthma, a chronic respiratory disorder marked by airway inflammation and bronchoconstriction, leads to exacerbations and significantly affects the pediatric population worldwide. Early asthma diagnosis is critical for improving disease management, reducing severe symptoms and economic costs, and preventing irreversible airway damage [1, 2].

This study aims to predict asthma in preschool children by analyzing co-morbidities, environmental, and socio-economic factors. We introduce a simple predictive tool, the *Asthma Ruler Tool* (ART), and compare its performance with the Leicester tool [3], focusing on non-invasive predictors including demographic, perinatal data, respiratory symptoms, eczema, and family atopy history, suitable for primary care assessment.

The main limitation is the reliance on parent-reported data, potentially biasing predictions and lacking objective asthma outcome measures. In order to address this limitations the ART was developed.

2 Methodology

This study aimed at creating a predictive tool for asthma in preschoolers, focusing on non-invasive, easily obtainable predictors during initial checkups.

We propose using ART incorporating the Leicester tool's covariates and adding others like parental smoking, education, family asthma/rhinitis history, wheezing frequency, and rhinoconjunctivitis, aiming for improved early diagnosis and intervention in pediatric asthma.

Furthermore, we evaluated our tool's efficacy using traditional statistical models, machine learning techniques, and deep neural networks, comparing it to the Leicester tool.

2.1 Traditional Statistical Models

Binomial Generalized Linear Models

Binomial Generalized Linear Models (GLMs) are statistical tools used to model binary outcomes. They offer valuable insights when the relationship between predictor variables and the response is inherently non-linear. These models extend the linear modeling framework to accommodate non-normal distributions of the response variable [4,5].

2.2 Traditional Machine Learning Methods

Traditional machine learning methods, such as GLMnet, Random Forest (RF), Gradient Boosting Machine (GBM), and Support Vector Machine with Class Weights (SVM-CW), rely on manually crafted features and linear classifiers for prediction. These methods are known for their simplicity and low computational requirements compared to deep learning techniques.

GLMnet

The GLMnet package provides a framework for fitting generalized linear models and analogous models using penalized maximum likelihood. It efficiently computes the regularization path for the lasso or elastic-net penalties across a range of values for the regularization parameters, α and λ , facilitating model selection and complexity control [6].

Random Forest

The Random Forest is an ensemble learning technique used for classification. It constructs multiple decision trees during the training phase and determines the mode of the classifications made by individual trees as the output. This method is known for its robustness and ability to prevent overfitting through ensemble averaging [7]. In order to do this properly several hyperparameters are tuned,

the number of trees, the maximum and minimum depth and the dimension of the leaf split.

Gradient Boosting Model

The Gradient Boosting Model constructs an ensemble of weak prediction models, mainly decision trees, to improve predictive accuracy. The method achieves remarkable flexibility by iteratively optimizing a differentiable loss function, making it suitable for a wide range of problems, including regression, classification, and ranking [8]. The aforementioned flexibility is reached employing many hyperparameters: the number of trees, the maximum depth, the dimension of the leaf split and a scale parameter. Shrinkage is then used to prevent overfitting.

Support Vector Machines with Class Weights

The Support Vector Machines with Class Weights algorithm adjusts the penalty associated with different classes during the training process to account for class imbalances. This variant prioritizes the importance of classes, resulting in improved model performance on imbalanced datasets [9]. The hyperparameters used for this class of model are the weights' class and the kernel.

2.3 Deep Neural Networks

DNNs are a subset of machine learning algorithms that use multiple layers to abstract high-level features from input data. They are useful for complex classification tasks that traditional statistical or machine learning methods cannot handle. Our study utilizes H2O's Deep Learning model, which is a multi-layer feedforward artificial neural network trained via stochastic gradient descent with back-propagation, to identify patterns that indicate asthma in the subject population [10,11]. The implemented DNNs has 2 hidden layers, a mini batch size of 16 and dropout as regularization technique.

3 Data Description

The study data set was obtained by distributing anamnestic questionnaires to the parents of the participating children. Not all of the original data set was used; in fact, only the children between the ages of 1 and 5 were selected for analysis. Table 1 provides full details of the data used in the research. For further analyses, the data set was divided into training and test sets using a 70/30 ratio, taking into account the proportion of asthmatic children. To ensure accurate and appropriate evaluation, it was also necessary to recode factor covariates using one-hot encoding. For all methods used, only the training set was considered for parameters tuning the overall parameters using 10-fold cross-validation procedures. Finally, the test set was used to validate the results.

Statistical analyses were performed using R software (v. 4.3). Statistical significance was defined as a p-value of ≤ 0.05 .

Table 1. Characteristics of the whole dataset.

	Asthma No N = 401	Asthma Yes N = 491	All N = 892	p-value
Age ($\pm$ SD) (years)	3.42 ($\pm$ 1.31)	3.84 ($\pm$ 1.17)	3.65 ($\pm$ 1.25)	< 0.001
Gender (%)				0.927
Female	172 (42.89%)	208 (42.36%)	380 (42.60%)	
Male	229 (57.11%)	283 (57.64%)	512 (57.40%)	
Height ($\pm$ SD) (cm)	98.39 ($\pm$ 11.77)	101.90 ($\pm$ 9.96)	100.32 ($\pm$ 10.95)	< 0.001
Weight ($\pm$ SD) (kg)	15.98 ($\pm$ 4.55)	17.38 ($\pm$ 4.16)	16.75 ($\pm$ 4.39)	< 0.001
bmi ($\pm$ SD) (kg/m ²)	16.33 ($\pm$ 2.11)	16.57 ($\pm$ 2.17)	16.46 ($\pm$ 2.14)	0.100
Traffic level (%)				0.009
No traffic	77 (19.20%)	113 (23.01%)	190 (21.30%)	
Little traffic	141 (35.16%)	157 (31.98%)	298 (33.41%)	
Moderate traffic	99 (24.69%)	86 (17.52%)	185 (20.74%)	
Heavy traffic	84 (20.95%)	135 (27.49%)	219 (24.55%)	
Distance to traffic (%)				0.454
<50metri	133 (33.17%)	163 (33.20%)	296 (33.18%)	
50=<100metri	61 (15.21%)	59 (12.02%)	120 (13.45%)	
100=<200metri	97 (24.19%)	117 (23.83%)	214 (23.99%)	
>=200metri	110 (27.43%)	152 (30.96%)	262 (29.37%)	
Mathernal smoking pregnancy (%)	39 (9.73%)	57 (11.61%)	96 (10.76%)	0.427
Father smoking pregnancy (%)	131 (32.67%)	169 (34.42%)	300 (33.63%)	0.632
Mathernal smoking early life (%)	37 (9.23%)	64 (13.03%)	101 (11.32%)	0.093
Father smoking early life (%)	120 (29.93%)	166 (33.81%)	286 (32.06%)	0.244
Relatives smoking early life (%)	102 (25.44%)	124 (25.25%)	226 (25.34%)	1.000
Father smoking now (%)	126 (31.42%)	166 (33.81%)	292 (32.74%)	0.494
Mother smoking now (%)	79 (19.70%)	124 (25.25%)	203 (22.76%)	0.059
Father's years smoking ($\pm$ SD) (years)	2.66 ($\pm$ 5.83)	3.61 ($\pm$ 6.82)	3.18 ($\pm$ 6.41)	0.026
Mother's years smoking ($\pm$ SD) (years)	5.85 ($\pm$ 8.56)	6.40 ($\pm$ 9.08)	6.15 ($\pm$ 8.85)	0.355
Moisture at home early life (%)	111 (27.68%)	153 (31.16%)	264 (29.60%)	0.290
Moisture at now (%)	90 (22.44%)	123 (25.05%)	213 (23.88%)	0.407
Day care attendance (%)	233 (58.10%)	207 (42.16%)	440 (49.33%)	< 0.001
Age of day care ($\pm$ SD) (years)	10.55 ($\pm$ 10.21)	7.38 ($\pm$ 10.22)	8.81 ($\pm$ 10.33)	< 0.001
Mother's age at birth ($\pm$ SD) (years)	32.26 ($\pm$ 4.82)	32.36 ($\pm$ 5.14)	32.31 ($\pm$ 5.00)	0.759
Father's educational level (%)				0.035
Elementary/None	7 (1.75%)	3 (0.61%)	10 (1.12%)	
Lower average	81 (20.20%)	130 (26.48%)	211 (23.65%)	
Upper average	205 (51.12%)	251 (51.12%)	456 (51.12%)	
Degree/Higher	108 (26.93%)	107 (21.79%)	215 (24.10%)	
Mother's educational level (%)				0.351
Elementary/None	5 (1.25%)	4 (0.81%)	9 (1.01%)	
Lower average	64 (15.96%)	100 (20.37%)	164 (18.39%)	
Upper average	206 (51.37%)	238 (48.47%)	444 (49.78%)	
Degree/Higher	126 (31.42%)	149 (30.35%)	275 (30.83%)	
Father's asthma (%)	46 (11.47%)	84 (17.11%)	130 (14.57%)	0.023
Father's rhinitis (%)	90 (22.44%)	153 (31.16%)	243 (27.24%)	0.005
Mother's asthma (%)	56 (13.97%)	91 (18.53%)	147 (16.48%)	0.082
Mother's rhinitis (%)	107 (26.68%)	145 (29.53%)	252 (28.25%)	0.387
Household members ($\pm$ SD)	3.77 ($\pm$ 0.79)	3.73 ($\pm$ 0.67)	3.75 ($\pm$ 0.73)	0.342

(continued)

Table 1. (continued)

Gestational age ($\pm$ SD) (years)	38.88 ($\pm$ 2.13)	38.91 ($\pm$ 2.06)	38.90 ($\pm$ 2.09)	0.836
Birth weight ($\pm$ SD) (grams)	3168.16 ($\pm$ 532.40)	3215.18 ($\pm$ 555.82)	3194.04 ($\pm$ 545.62)	0.199
Delivery method (%)				0.093
Caesarean	239 (59.60%)	264 (53.77%)	503 (56.39%)	
Eutocytic	162 (40.40%)	227 (46.23%)	389 (43.61%)	
Birth distress (%)	42 (10.47%)	47 (9.57%)	89 (9.98%)	0.738
Low flow oxygen (%)	28 (6.98%)	29 (5.91%)	57 (6.39%)	0.606
Days on low flow oxygen ($\pm$ SD) (days)	0.46 ($\pm$ 3.13)	0.19 ($\pm$ 1.27)	0.31 ($\pm$ 2.30)	0.107
NIC (%)	30 (7.48%)	25 (5.09%)	55 (6.17%)	0.182
Days of NIC ($\pm$ SD) (days)	0.72 ($\pm$ 6.37)	0.27 ($\pm$ 1.69)	0.47 ($\pm$ 4.45)	0.170
Breastfeeding (%)	254 (63.34%)	281 (57.23%)	535 (59.98%)	0.074
Breastfeeding duration ($\pm$ SD) (months)	10.97 ($\pm$ 14.25)	7.28 ($\pm$ 11.01)	8.93 ($\pm$ 12.70)	<0.001
Wheezing episodes in the last 12 months (%)	111 (27.68%)	338 (68.84%)	449 (50.34%)	<0.001
Atopic dermatitis (%)	104 (25.94%)	114 (23.22%)	218 (24.44%)	0.389
Rhinoconjunctivitis (%)	97 (24.19%)	208 (42.36%)	305 (34.19%)	<0.001
Eczema (%)	10 (2.49%)	6 (1.22%)	16 (1.79%)	0.242
Rhinitis (%)	54 (13.47%)	147 (29.94%)	201 (22.53%)	<0.001
Pets in early life (%)	59 (14.71%)	53 (10.79%)	112 (12.56%)	0.098
Pets now (%)	67 (16.71%)	69 (14.05%)	136 (15.25%)	0.315
Occurrence of wheezing in life (%)	69 (17.21%)	252 (51.32%)	321 (35.99%)	<0.001
Lower respiratory tract symptoms (%)	116 (28.93%)	223 (45.42%)	339 (38.00%)	<0.001
Upper respiratory tract symptoms (%)	192 (47.88%)	237 (48.27%)	429 (48.09%)	0.962

4 Results

This section presents the findings of our study, which includes three primary categories of predictive models: traditional statistical models, traditional machine learning methods, and deep neural networks. We evaluated their performance using two key metrics:

- **Accuracy:** The overall correctness of predictions.
- **AUC (Area Under the ROC Curve):** A measure of the model’s ability to distinguish between positive and negative instances across different probability thresholds.

Variable selection is a critical factor in asthma prediction. While some features contribute significantly to accurate prediction, others may introduce noise or confounding effects. Our study highlights the importance of careful covariate selection to improve prediction accuracy.

Table 2 shows the results on the test set of both the ART and the Leicester tool, which was used as a benchmark. The metrics used to evaluate the performance of these models are accuracy and AUC. The results show that the

ART outperforms the Leicester tool on both metrics across all methods, indicating robust predictive capabilities. DNNs achieves the best results: Accuracy of 0.7428 and AUC of 0.790, demonstrating the potential of deep learning for predictive modeling.

The addition of covariates-parental smoking history, education, family history of asthma/rhinitis, wheeze episodes, and rhinoconjunctivitis-enhanced the dataset and aided in the non-invasive diagnosis of asthma in preschool children. Overall, ART’s superior performance across methods highlights its effectiveness and versatility in predictive modeling, with unmatched accuracy and AUC.

Table 2. Comparative table of results obtained in the test set for the Asthma Ruler Tool and the Leicester Tool

	Asthma Ruler Tool		Leicester Tool	
	Accuracy	AUC	Accuracy	AUC
Binomial GLM	0.7138	0.747	0.6211	0.699
GLMnet	0.7286	0.784	0.5948	0.639
RF	0.7398	0.788	0.6283	0.661
GBM	0.7212	0.788	0.6171	0.654
SVM-CW	0.7398	0.786	0.6059	0.662
DNNs	0.7428	0.790	0.5399	0.539

5 Discussion

Creating a definitive non-invasive tool to predict asthma in children remains a challenge. This study aims to develop such a tool. Our proposal outperforms the Leicester tool in all models and metrics provided. This difference in performance could be attributed to variations in model implementation, data used, or parameter settings. Further investigation into these aspects could provide valuable insights and potentially lead to improvements in the tools.

This analysis offers a valuable comparison of the Asthma Ruler Tool and Leicester tools using various machine learning models. However, it also emphasizes the significance of comprehensive performance evaluation and the necessity for further investigation to fully comprehend the observed differences in performance.

6 Conclusion

Our work provides a promising starting point for developing a non-invasive tool to predict asthma in preschool children. However, there is room for improvement. Enhancing the Leicester model by incorporating additional variables could lead to a more comprehensive and accurate prediction, paving the way for future research in this field.

References

1. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (2020). www.ginasthma.org
2. Bahadori, K., Doyle-Waters, M.M., Marra, C., et al.: Economic burden of asthma: a systematic review. *BMC Pulm. Med.* **9**, 24 (2009). <https://doi.org/10.1186/1471-2466-9-24>
3. Pescatore, A.M., Dogaru, C.M., Duembgen, L., Silverman, M., Gaillard, E.A., Spycher, B.D., Kuehni, C.E.: A simple asthma prediction tool for preschool children with wheeze or cough. *J. Allergy Clin. Immunol.* **133**(1), 111-8.e1-13 (2014). <https://doi.org/10.1016/j.jaci.2013.06.002>.
4. McCullagh, P., Nelder, J.A.: *Generalized Linear Models* (2nd ed.). Routledge (1989). <https://doi.org/10.1201/9780203753736>
5. Dunn, P.K., Smyth, G.K.: *Generalized Linear Models With Examples in R*. Springer Texts in Statistics. Springer, New York, NY (2018). https://doi.org/10.1007/978-1-4419-0118-7_9
6. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**(1), 1–22 (2010). <https://doi.org/10.18637/jss.v033.i01>
7. Breiman, L.: Random Forest. *Mach. Learn.* **45**(1), 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232 (2001). <https://doi.org/10.1214/aos/1013203451>
9. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 1–27 (2011). <https://doi.org/10.1145/1961189.1961199>
10. Wright, M.N., Ziegler, A.: ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**(1), 1–17 (2017). <https://doi.org/10.18637/jss.v077.i01>
11. Candel, A., Parmar, V., LeDell, E., Arora, A.: *Deep learning with H2O*. H2O.ai, Inc., Mountain View, CA, USA (2016). <https://h2o.ai/resources/>



The Integration Paths of Refugees and Asylum Seekers in Italy

Camilla Pangallo¹(✉), Cinzia Conti², Fabio Massimo Rottino²,
and Oliviero Casacchia¹

¹ Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy
{camilla.pangallo,oliviero.casacchia}@uniroma1.it

² Italian National Institute of Statistics (ISTAT), Via Cesare Balbo 16, 00184 Rome, Italy
{ciconti,rottino}@istat.it

Abstract. In the last decade, Italy has been a country characterised by the arrival of significant migratory flows for humanitarian reasons. For many years, since the refugee crisis in the Mediterranean in 2015, the presence of asylum seekers and refugees in our country has been considered as transitory but today it is evident that this group of migrants has begun to stabilise in Italy through integration processes. It is therefore relevant to focus on the integration paths of refugees and asylum seekers, who present different characteristics from other migrants. The main objective of this work is to identify the profile and main demographic characteristics of regular asylum seekers who *stabilise* in Italy because they are present in the territory six years after the issue of their residence permit. The study is based on data from residence permits and a longitudinal cohort approach was used applying record linkage techniques. Finally, a logistic model was applied to the data.

Keywords: Refugees · Asylum seekers · Integration paths · Longitudinal cohort approach

1 Introduction

Today, many countries are more committed to addressing global economic imbalances, divergent demographic trends and climate change than to migration, which will become a necessity in the decades to come for countries at all income levels. If managed well, migration can be a force for prosperity and can contribute to achieving the UN Sustainable Development Goals [6]. About 184 million people, or 2.3% of the world's population, live outside their country of origin and almost half of them are in low- and middle-income countries [6].

Due to new conflicts and the continuation of decades-long crisis situations globally, the number of forcibly displaced persons worldwide has been estimated at 108.4 million at the end of 2022. This figure includes 35.3 million refugees, 5.4 million asylum seekers, 62.5 million internally displaced persons and 5.2 million other persons in need

of international protection [3]. In particular, the number of refugees and internally displaced persons has grown exponentially, reaching around 65 million at the end of 2015, mostly hosted in countries neighbouring conflict zones. This international situation is also reflected in Europe, albeit with smaller numbers [4].

Italy has also seen a significant increase in international protection migrants since 2013, peaking in 2017, following the so-called Mediterranean crisis. In 2022, 105,129 people reached Italy through landings, marking an increase of 55.79% compared to the previous year. In terms of migratory routes, 53,310 came from Libya, 32,371 from Tunisia and 16,205 from Turkey; about 13,000 people arrived by the Balkan Route at the land border in Trieste [2]. Many people fleeing Ukraine also reached Italy in 2022. In fact, between 8 March and 31 December, 151,435 beneficiaries of temporary protection were registered in the country [2].

Integration is therefore a fundamental area of the asylum system. Too often, for years, European countries, including Italy, have approached the arrival of refugees and asylum seekers as an emergency, implementing solutions such as relocation, resettlement and the outsourcing of EU border control, rather than attempting to implement structural policies capable of addressing the situation in a long-term perspective [1].

Local integration therefore helps to ensure that refugees can build a new life for themselves in host countries. Through this process it is possible to create an environment of long-term stability, enabling refugees to rebuild their lives, to actively participate in the economic and social life of the host country and to support themselves and their families [5].

2 Objectives

The preliminary objective of the work is to better understand and describe the flows of regular immigrants who have arrived in Italy in the last decade, in particular by analysing the number of new arrivals according to the reason for their residence permit.

The main objective of the study is to analyse in more detail migrants coming to Italy for international protection reasons, given the sharp increase in these flows in recent years. Within this group, migrants who have come to Italy to apply for asylum represent the most significant and consistent category. It is therefore no longer possible to speak of the temporary presence of asylum seekers in Italy because they represent a community that over time is trying to settle and integrate in our country. For this reason, the analysis aims to identify the profile and main demographic characteristics of regular asylum seekers who remain in Italy six years after their permit was issued, of those migrants we can define as *stayers*.

In this study, we propose an updated analysis [3] considering the latest available data (stock data until 1 January 2023 and flow data until 2022).

3 Data and Methods

The study is based on data from residence permits issued to non-EU citizens, particularly those permits for reasons related to political asylum applications. This information is provided by the Ministry of the Interior to Istat, which is responsible for harmonising and processing this data.

The data on residence permits, albeit in a different way from asylum applications and with certain limitations¹, can provide useful information on the flows and presence of people in our country for asylum or other forms of international protection.

In this work, stock data from 2011 to 2023 and flow data from 2011 to 2022 were used.

The methodological novelty in dealing with these data on residence permits is the use of a *longitudinal cohort approach*. The data for the various years were linked through a deterministic record linkage, using unique identification codes. Through the linkage technique it is possible to follow individuals over time and to verify the continuity of their regular presence on the territory. However, with this approach it is only possible to monitor regular migrants and if the person does not renew their residence permit but remains in Italy, they “disappear” from the residence permit dataset. Good quality analyses can still be performed as the unique code is available for about 90% of the cases.

A logistic regression analysis was conducted on asylum seekers who entered in 2007 and between 2011 and 2017; in addition, each year of entry was linked with the stock six years later. The following variables were used in the model:

- the response variable provides information on the presence or absence of the asylum seeker on Italian territory six years after the issue of the residence permit. Thus, there are two categories for this response variable:
 - *stayers* are the asylum seekers who arrived in Italy and who are still present in our country six years later;
 - *no longer present* are asylum seekers who arrived in Italy and are no longer present in our country 6 years after the issue of the residence permit;
- the five covariates used in the analysis are: gender; citizenship (the first five citizenships plus the sixth category “other nationality”); age (15–20 years, 21–25 years, 26–35 years, 36 years and over); Italian territorial units (five macro-regions: North-West, North-East, Centre, South, Islands); and the issue year of the first residence permit (2007, 2011, 2012, 2013, 2014, 2015, 2016, 2017).

For the model, the proc logistic of the SAS software was used, with a recursive procedure, using the backward elimination selection technique. In the present case it was not necessary to remove any variables.

4 Results

The analysis of permit data gives us a picture of the type of flows and the presence of migrants in our country for the reason of the permit. A total of 3,074,746 permits were issued in Italy between 2011 and 2022, of which 23.4% were for asylum (Fig. 1).

¹ The data of residence permits for “asylum seekers” register the arrival of migrants later than for asylum applications due to the time lag between the application and the issuance of the permit. In addition, in some provinces the registration of asylum seekers’ permits is affected by undercounting.

In the last decade there has been an unprecedented contraction of flows for work reasons and a substantial stability of those related to family reasons (with the exception of 2020), considering that the latter flows are the most consistent.

Analysing the flows of migrants arriving in Italy for international protection, we observe that the number of asylum seekers entering in Italy has experienced various fluctuations over the years. After an increase in 2014, there was a significant peak in arrivals in 2017. The Covid pandemic then led to a drastic decrease in the overall number of arrivals. After the crisis due to the pandemic, we can observe an increase. During 2022 the flows of migrants arriving from Ukraine were noticed, but during 2022 other asylum seekers from other countries arrived in Italy. Above all, there were Bangladeshis (9,616 residence permits) and Pakistanis (8,396 documents). In third place were Egyptians with almost 5,000 permits for reasons related to asylum. These three citizenships, and Ukrainians covered approximately 53% of permits issued for reasons of asylum or international protection in 2022. In 2022 the Nigerians were the fifth citizenship with 3,576 permits.

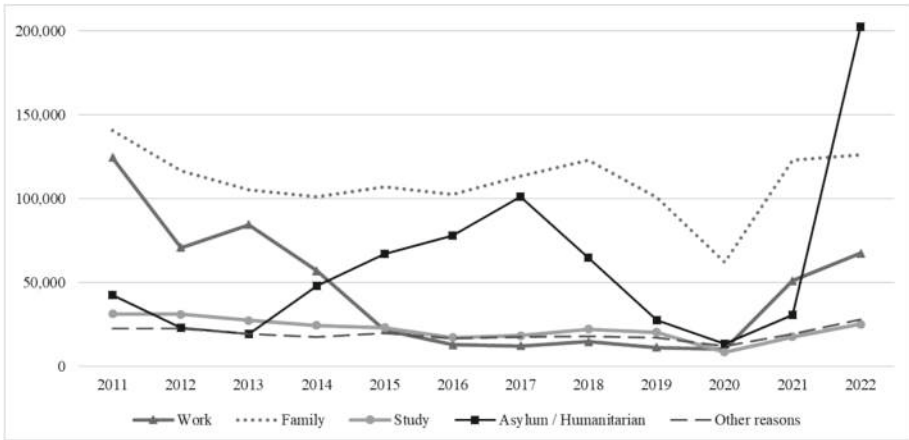


Fig. 1. New permits issued during the reference year by reason, 2011–2022, absolute values. **Source:** Elaboration on Istat data.

The analysis focuses on asylum seekers in Italy using a longitudinal approach. Specifically, we aim to identify the profile and demographic characteristics of *stayers*, asylum seekers who have settled in Italy and remain in the country six years after receiving their residence permit.

The results were obtained by applying the logistic regression model to the data and the estimates were significant ($p\text{-value} < 0.05$).

As shown in Fig. 2, it is relevant to note that women are more likely than men to settle in our country; in fact, women are 55% more likely than men to remain in Italy. Moreover, among asylum seekers, it is noticeable that Senegalese citizens are considered the least likely to remain in Italy. On the contrary, those with Bangladeshi (+85%) and Pakistani (+75%) nationalities show higher propensities to leave Italian territory during the period examined.

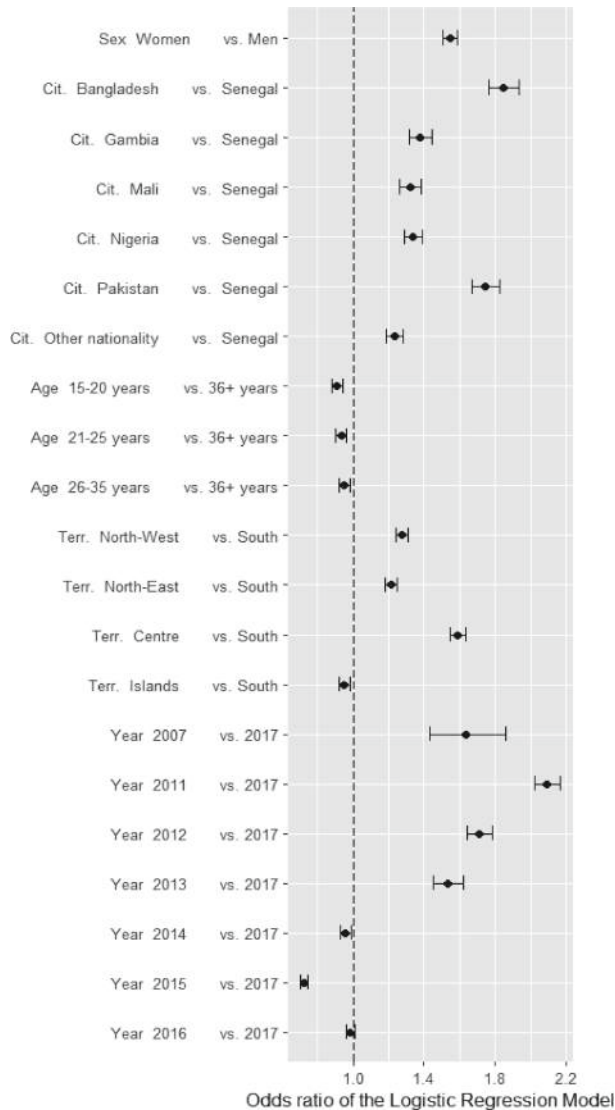


Fig. 2. Results of the logistic regression model: some characteristics of *stayers* who arrived in Italy. **Source:** Elaboration on Istat data.

Considering age, asylum seekers under the age of 36, particularly the youngest in the 15–20 age group, tend to stay in Italy for shorter periods than migrants of an older age. The territorial distribution was taken into account by considering the geographical location of the prefecture that issued the residence permit. Asylum seekers living in Central Italy are 59% more likely to stay permanently than asylum seekers living in the South, while those living on the Islands tend to stay in the country only temporarily (the odds ratio value is 0.95). Finally, when examining the year of arrival of asylum seekers

in Italy, it is evident that asylum seekers who entered in 2011 are twice as likely to settle in Italy as flows arriving in 2017. This propensity is also observed for the 2007, 2012 and 2013 flows, but with a lower intensity. On the other hand, it is observed that for the 2014, 2015 and 2016 flows there is a lower probability of migrants staying in Italy than for the 2017 flow, with this trend being more evident for the 2015 arrivals (the odds ratio value is 0.72).

5 Conclusions

In this paper, we started with a brief analysis of regular new arrivals by type of permit issued. From this study, it became clear that in Italy, in recent years, there has been periods of consistent increase in arrivals related to requests for international protection, with the largest group of these migrants being political asylum seekers. Consequently, the study continues by focusing more closely on the group of asylum seekers in order to understand the main characteristics of those who tend to stabilise on Italian territory.

The work showed that among asylum seekers, migrant women are more likely to settle in Italy than men, those with Bangladeshi and Pakistani citizenship than Senegalese, and those who are older (over 36 years) than younger. But also those who settle in the central regions of Italy compared to the south of Italy and migrants who arrived in our country in 2011, 2012 and 2013 compared to the flows of asylum seekers who arrived in 2017. The results obtained in this work with the most up-to-date data confirm the evidence found in our other work [3].

It is therefore crucial to understand the life paths of asylum seekers because today they can no longer be addressed as an emergency situation but represent a reality that is increasingly consolidating in the territory through integration processes. Therefore, the successful integration of this group brings social, cultural and economic benefits to the country and the host population.

References

1. Archibugi, D., Cellini, M., Vitiello, M.: Refugees in the European Union: from emergency alarmism to common management. *J. Contemp. Eur. Stud.* **30**(3), 487–505 (2022)
2. ASGI: Country Report Italy, 2022 Update (2023)
3. Casacchia, O., Conti, C., Pangallo, C., Rottino, F. M.: Beyond the emergency: characteristics and behaviour of refugees and asylum seekers in Italy. In: *RIVISTA ITALIANA DI ECONOMIA, DEMOGRAFIA E STATISTICA*. – ISSN 0035–6832. – Volume LXXVII: n.3 Luglio-Settembre 2023, pp. 115–126 (2024)
4. UNHCR: Focus group sul tema dell'integrazione report finale (2017)
5. UNHCR: Global Trends. Forced Displacement in 2022 (2023)
6. World Bank: World Development Report 2023: Migrants, Refugees, and Societies. World Bank, Washington, DC (2023)



Species Distribution Modeling Approach for Biased Citizen Science Data

Greta Panunzi¹(✉), Jafet Belmont², Janine Illian², and Sara Martino³

¹ Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy

greta.panunzi@uniroma1.it

² School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

³ Department of Mathematical Sciences, Trondheim, Norway

Abstract. Ecological research emphasizes the criticality of comprehending species distribution for the formulation of efficient conservation strategies. The advent of Citizen Science initiatives has transformed the landscape of species occurrence data collection, offering a cost-efficient avenue for monitoring wildlife across diverse spatiotemporal scales. Nevertheless, the absence of standardized sampling protocols within these initiatives poses analytical hurdles, leading to skewed sampling efforts biased towards easily accessible or extensively studied regions. This study analyzes the national butterfly monitoring program in the UK. It uses a point process framework that combines spatial and temporal covariates to examine spatial and temporal trends of butterfly species occurrences. The methodology merges two critical elements: (i) the geographical locations visited by participating volunteers and (ii) the presence records of the species under investigation, utilizing the INLA/inlabru framework. The study highlights the need to reduce sampling bias to improve the accuracy and reliability of species distribution modeling. By combining the benefits of Citizen Science data and robust modeling methodologies, this approach provides a more nuanced understanding of species distribution dynamics. It can thus help strengthen the basis for more effective conservation efforts.

Keywords: Citizen Science · INLA · Integrated Models · Occupancy · Preferential Sampling

1 Introduction

A primary goal in ecological studies is to understand the geographic and environmental factors influencing species occurrences [Elith and Leathwick, 2009]. Traditional observation protocols, however, are highly costly and time-consuming. Citizen Science (CS) monitoring programs have become essential for collecting large volumes of species distribution data at low cost across broad spatiotemporal scales, facilitated by advancements in technology and information systems [Theobald et al., 2015]. Over the last decade, CS data have proven valuable for monitoring wildlife populations, leading to an increase in studies utilizing

such data for species distribution modeling [Bonney et al., 2009, 2014, Theobald et al., 2015, Feldman et al., 2021]. However, analyzing CS data remains challenging due to the lack of standardized sampling protocols, resulting in uneven sampling efforts over time and space, often biased towards human activity centers or easily accessible locations [Martin et al., 2012]. Data integration poses another significant challenge in CS, as diverse methods used by volunteers lead to heterogeneous and potentially inconsistent data [Bonney et al., 2009]. Discrepancies between CS data and professionally collected data are common, and combining different data sources is necessary to improve the accuracy and quality of species distribution predictions [Aceves-Bueno et al., 2017]. Yet, reliably combining these sources is challenging due to variations in design, spatial and temporal coverage, and potential sampling biases among different CS recording schemes. In this study, we consider a CS application related to the monitoring scheme of butterflies in the United Kingdom. Over the past few years, numerous European butterflies have seen a considerable decline [Van Swaay et al., 2006]. To obtain standardized information and early signals regarding the decline of butterflies, a national butterfly monitoring scheme was established in the UK in 1976 [Pollard, 1977, van Swaay et al., 2008]. Over time, this monitoring program has incorporated various observation methods, all based on observations from volunteers who observe differently depending on the observation program they choose to participate in. Working with CS data the first issue that needs to be addressed is that observers tend to visit areas where they expect to observe more individuals. We correct this bias estimating an effort intensity by utilizing information from all available observations in an attempt to construct an observation effort for our focal species. We utilize an advanced point process framework that integrates various data sources and incorporates spatial covariates, allowing for the analysis of spatial trends in butterfly species occurrences. The study underscores the importance of addressing sampling bias to improve the accuracy and reliability of species distribution modeling. By leveraging the synergies between Citizen Science data and robust modeling methodologies, this approach enables a deeper and more nuanced understanding of species distribution dynamics. Consequently, it establishes a strong foundation for more targeted and impactful conservation efforts.

2 Available Data

The data analyzed in this study encompasses two distinct survey schemes. The first scheme, known as *Wider Countryside Butterfly Survey* (referenced as WCBS [UK Butterfly Monitoring Scheme, 2021]), is a standardized monitoring program aimed at tracking widespread butterflies across the UK. Under this scheme, surveyors are tasked with conducting surveys in designated squares at least twice during the July and August period, with visits spaced ten days apart. The second scheme, *New Millennium* (referenced as NM [Butterfly Conservation, 2022]), is a national recording initiative for butterflies, administered by Butterfly Conservation in collaboration with the Biological Records Centre (BRC). This dataset

comprises 400,000 distribution records of butterflies gathered by volunteers. The NM is an opportunistic survey without a standardized sampling protocol, i.e. biological records can be obtained from any user, in any location, at any time. In this study, we focus on the *Speckled Wood* species, which holds particular significance in environmental studies as one of the *Habitat specialists* species [UK Government, 2024]. As depicted in Fig. 1, the New Millennium dataset contains a considerably higher volume of data, despite with more variability. Hence, it becomes important to integrate these data with those from the WCBS scheme. To ensure comparability between the datasets, we restrict our analysis to observations recorded during the months of July and August.

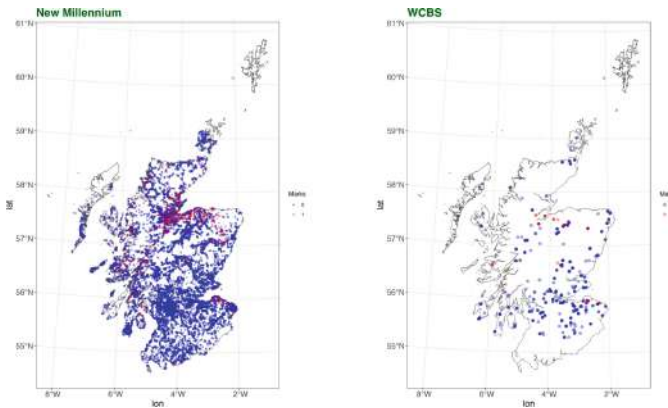


Fig. 1. Observations from the two datasets. The red dots represent observations of the species of interest, while the blue dots represent observations of other species.

3 Methods

A point process model was used to describe the locations where any species was observed. For each data source, it was assumed that the location of all butterflies in space ($s \in S \subset \mathbb{R}^2$) was properly described by a point process whose intensity function $\lambda_i(s)$ (when $i = 1$, it signifies NM data, while $i = 2$ denotes WCBS data) is additive on the log-scale:

$$\log(\lambda_i(s)) = \mathbf{X}^T(s)\beta_i + \omega_i(s) \quad i = 1, 2 \quad (1)$$

where $\mathbf{X}(s)$ is a set of covariates detected at location s with linear effects β to be estimated, and $\omega_i(s)$ is the spatial process capturing the spatial structure of all the locations where at least one species has been observed.

The distribution of the species of interest is also modeled as a point process with log intensity:

$$\log(\alpha_i(s)) = \mathbf{Y}^T(s)\gamma_i + \delta_i(s) + \xi_i\omega_i(s) \quad i = 1, 2 \quad (2)$$

where $\mathbf{Y}(s)$ is a set of covariates detected at location s with linear effects γ to be estimated, $\omega_i(s)$ is the spatial process in Eq. 1, ξ_i is a linear term to be estimated and $\delta_j(s)$ is the species-specific spatial process. Although it would have been theoretically possible to consider a complex spatio-temporal model [Yuan et al., 2017], $\omega(s, t)$, the limited number of sightings each year for the WCBS dataset did not provide enough information to allow robust estimates. Therefore we chose to estimate a purely spatial model. This model was initially fitted independently to each data source j and then fitted jointly producing an integrated model where the parameters δ_j and β_j were shared. Then, the joint model was defined as follows:

$$\begin{aligned} \log(\lambda_1(s)) &= \mathbf{X}^T(s)\beta_1 + \omega_1(s) \\ \log(\lambda_2(s)) &= \mathbf{X}^T(s)\beta_2 + \omega_2(s) \\ \log(\alpha_1(s)) &= \mathbf{Y}^T(s)\gamma_1 + \xi_1\omega_1(s) + \delta(s) \\ \log(\alpha_2(s)) &= \mathbf{Y}^T(s)\gamma_2 + \xi_2\omega_2(s) + \delta(s) \end{aligned} \quad (3)$$

The choice to consider a joint model is based on the idea that the specific Gaussian process of the observation process depends on the scheme we are considering, therefore it is specific to the dataset we refer to. Conversely, the Gaussian process of the species intensity is a characteristic of the species we are studying, not of how we observe the data, therefore it must be the same for both observation collection methods.

To implement the model in a Bayesian framework, we must define prior distributions for all model parameters. We chose a non-informative Gaussian prior for β with mean 0 and variance 100. For the parameters in the spatial fields ω and δ in Eqs. 1, 2 and 3, we used Penalized-complexity (PC) priors [Fuglstad et al., 2019] with $P(\rho < 300) = 0.5$ and $P(\sigma > 1) = 0.5$. Finally, for ξ , we chose a normal prior with a mean of 0.5 and a variance of 0.3. To estimate Gaussian fields (GFs) we considered the SPDE (Stochastic Partial Differential Equation) representation, pioneered by Lindgren et al. [2011], which offers numerous computational advantages. This approach enables efficient computation and inference on spatial models, making it particularly suitable for managing large datasets and intricate spatial structures. The SPDE method relies on a mesh to discretize the spatial domain. Unlike grid discretization, this approach eliminates the need to approximate the locations of points (such as grid cell centroids); instead, exact occurrence coordinates can be directly considered.

4 Preliminary Results

Looking at the preliminary results in Fig. 2, the difference in estimations becomes apparent when considering the two different datasets. With the NM dataset, there is a significantly high intensity in the northern region of Scotland, while

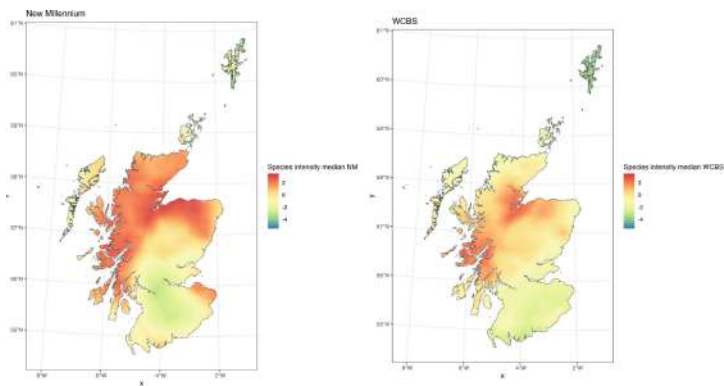


Fig. 2. The posterior estimates of the median of the log intensity of the *Speckled Wood* occurrence obtained considering two datasets separately (Eq. 2).

with the WCBS dataset, the intensity is higher near the *Clyde Muirshiel Regional Park* and near the *Inverness area*.

The diversity of estimates confirms the importance of using both data sources. In Fig. 3, the median intensity estimate obtained from considering the joint model of the system of Eq. 3 is depicted. As expected, combining information from the two observation schemes yields different estimates. Specifically, compared to estimates obtained solely from NM data, the intensity is lower, whereas when compared to WCBS data, the intensity estimate increases in the northern region of Scotland.

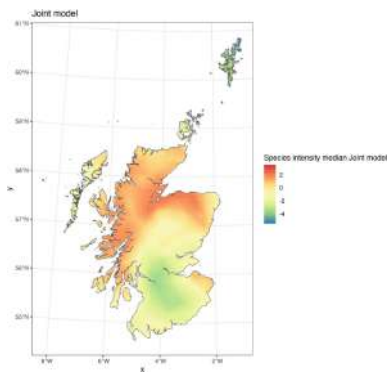


Fig. 3. The posterior estimates of the median of the log intensity of the *Speckled Wood* occurrence considering the joint model.

Finally, by examining the parameters ξ_1 and ξ_2 in Table 1, it is possible to observe how the Gaussian field of the observation process indeed influences the

Table 1. Summary of key results from estimating 3.

Parameter	Mean $\pm$ sd	95% CI
ω_1	495.826 ± 103.017	[331.627, 735.13]
ω_2	246.657 ± 36.039	[160.264, 269.887]
δ	244.243 ± 47.354	[167.377, 353.01]
ξ_1	0.862 ± 0.101	[0.663, 0.910]
ξ_2	0.749 ± 0.102	[0.548, 0.951]

estimation of the species intensity. This confirms the importance of including effort estimation, as the observation process significantly influences the estimation of species intensity.

5 Conclusions

This study underscores the critical importance of understanding species distribution dynamics for the development of effective conservation strategies within ecological research. Our analysis focused on the national butterfly monitoring program in the UK, utilizing a point process framework that integrates different data sources to explore spatial trends of butterfly species occurrences. Despite the challenges posed by heterogeneous data sources and potential biases, our study showcases the potential of integrating diverse datasets to improve the accuracy and reliability of species distribution modeling. Correcting intensity estimates for sampling bias is crucial. By leveraging information from other species, we attempted to estimate the sampling effort, which is not always available in Citizen Science data. Looking ahead, continued efforts to standardize sampling protocols within Citizen Science initiatives will be crucial for enhancing the quality and consistency of data collection. Additionally, further developments in modeling techniques and computational methods will enable more comprehensive analyses of species distribution patterns, paving the way for more effective conservation strategies and the preservation of biodiversity for future generations.

References

Aceves-Bueno, E., et al.: The accuracy of citizen science data: a quantitative review. *Bullet. Ecol. Soc. Am.* **98**(4), 278–290 (2017)

Bonney, R., et al.: Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience* **59**(11), 977–984 (2009)

Bonney, R., et al.: Next steps for citizen science. *Science* **343**(6178), 1436–1437 (2014)

Butterfly Conservation. Butterfly distributions for Scotland from butterfly conservation and the biological records centre (2022). <https://doi.org/10.15468/tqf8z3>. Accessed 04 Mar 2024

- Elith, J., Leathwick, J.R.: Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **40**, 677–697 (2009)
- Feldman, M.J., Imbeau, L., Marchand, P., Mazerolle, M.J., Darveau, M., Fenton, N.J.: Trends and gaps in the use of citizen science derived data as input for species distribution models: a quantitative review. *PLoS One*, **16**(3), e0234587 (2021)
- Fuglstad, G.-A., Simpson, D., Lindgren, F., Rue, H.: Constructing priors that penalize the complexity of gaussian random fields. *J. Am. Stat. Assoc.* **114**(525), 445–452 (2019)
- Lindgren, F., Rue, H., Lindström, J.: An explicit link between gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc.: Ser. B (Statist. Methodol.)* **73**(4), 423–498 (2011)
- Martin, L.J., Blossey, B., Ellis, E.: Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. *Front. Ecol. Environ.* **10**(4), 195–201 (2012)
- Pollard, E.: A method for assessing changes in the abundance of butterflies. *Biol. Cons.* **12**(2), 115–134 (1977)
- Theobald, E.J., et al.: Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biol. Conserv.* **181**, 236–244 (2015)
- UK Butterfly Monitoring Scheme. Wider Countryside Butterfly Scheme (WCBS) (2021). <https://doi.org/10.15468/jgm7ci>
- UK Government. Butterflies in the United Kingdom: habitat specialists and species of the wider countryside 1976 to 2021. Official Statistics (2024). <https://www.gov.uk/government/statistics/butterflies-in-the-wider-countryside-uk/butterflies-in-the-united-kingdom-habitat-specialists-and-species-of-the-wider-countryside-1976-to-2021#habitat-specialists>. Accessed 04 Mar 2024
- Van Swaay, C., Warren, M., Loïs, G.: Biotope use and trends of european butterflies. *J. Insect Conserv.* **10**, 189–209 (2006)
- van Swaay, C.A.M., Nowicki, P., Settele, J., Van Strien, A.J.: Butterfly monitoring in Europe: methods, applications and perspectives. *Biodivers. Conserv.* **17**, 3455–3469 (2008)
- Yuan, Y., et al.: Point process models for spatio-temporal distance Martindata from a large-scale survey of blue whales (2017)



A Nonparametric Approach to Model Event-Data on Linear Networks

Simone Panzeri¹(✉), Aldo Clemente¹, Eleonora Arnone²,
and Laura M. Sangalli¹

¹ MOX – Dipartimento di Matematica, Politecnico di Milano, P.zza L. Da Vinci 32,
Milan 20133, Italy

{simone.panzeri, aldo.clemente, laura.sangalli}@polimi.it

² Dipartimento di Management, Università degli Studi di Torino, Corso Unione
Sovietica 218 bis, Torino 10134, Italy
eleonora.arnone@unito.it

Abstract. In this article, we focus on spatio-temporal point patterns observed on linear networks and recorded continuously over time. We present a nonparametric methodology for spatio-temporal intensity estimation in inhomogeneous Poisson point processes. The approach combines maximum likelihood estimation with roughness penalties based on differential operators in time and in space, the latter defined over the linear network domains. This balances data adaptation and smoothness of the estimate. We establish theoretical properties related to the proposed estimator. For the implementation, we rely on advanced techniques coming from optimization and numerical analysis. The discretization of the estimation problem combines finite elements in space, designed for linear networks, and B -splines in time, ensuring flexibility at feasible computational costs. We present an application to real data concerning road accidents occurred in Bergamo, Italy, in 2015-2022. This offers the opportunity to validate the proposed method, and to assess its performance in comparison with state-of-the-art techniques.

Keywords: Spatio-temporal intensity estimation · inhomogeneous Poisson point process · Linear networks · Penalized maximum likelihood estimation

1 Motivation

Various methodologies have been proposed to characterize the spatial distribution of point patterns, for instance by estimating the first-order intensity functions of the associated spatial Poisson point processes. Most of these techniques have focused on point patterns observed on standard planar domains. Only in recent years, some initial tools specifically designed for analyzing point patterns on linear network domains have been introduced. This has been motivated by the growing availability of geo-referenced data collected, for instance, by GPS systems, and related to human activities, occurring in network-constrained domains,

such as road networks and public transport networks. In particular, a wide range of applications to real data arises in traffic engineering and social sciences.

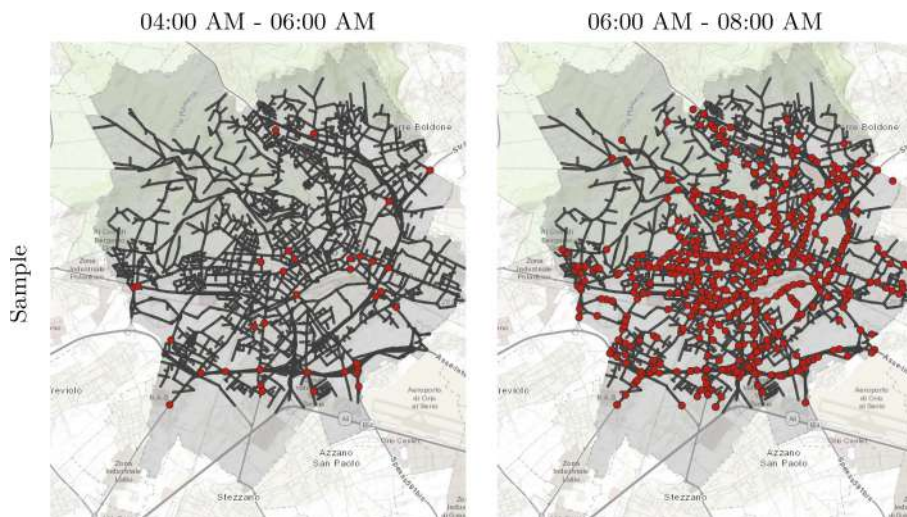


Fig. 1. Locations of road accidents occurred in Bergamo, Italy, from 2015 to 2022. The observed point pattern exhibits different behaviors over time across the road networks. Although the plots refer to time intervals of 2 h, the data are continuously collected over time.

Figure 1 illustrates an example of such a point pattern, displaying the locations of road accidents occurred in the city of Bergamo, Italy, from 2015 to 2022. The dataset is recorded by the municipal police and is publicly available at <https://www.dati.lombardia.it/> (last accessed 14 February 2024). The dataset comprises 13 033 road accidents that took place in the considered period. Besides exhibiting complex spatial dependencies over the road network domain of Bergamo, the evolution of occurrences strongly depends on time slots, with rush hours characterized by a greater risk of accidents due to congested traffic conditions.

Applied problems of this kind have played a role in stimulating research, directing it towards the idea that the physical domain can significantly influence and provide information about the phenomenon under study. In this regard, existing methods for point processes on $\mathbb{R}^2$ may not be appropriate, suggesting, as originally introduced by Okabe and Yamada (2001), that Euclidean distances should be replaced by shortest-path distances.

Intensity estimation in spatial Poisson point processes has been the subject of an extensive literature, with several techniques that have been introduced. The extension of these studies to the spatio-temporal framework has been thoroughly explored (for the theory of spatio-temporal Poisson point processes, the reader may refer to González et al., 2016; Diggle, 2013, and references therein), despite the inevitably increasing modeling complexity compared to the spatial case. This

complexity primarily arises from the inclusion of a further (temporal) dimension into the analysis and therefore implies non-negligible computational challenges. Nevertheless, some techniques to handle spatio-temporal Poisson point processes have recently been introduced, even in the case of occurrences evolving on linear networks over time. To address the intricate spatio-temporal dependencies characterizing the phenomenon under consideration, these techniques adopt different strategies. Some of them perform data aggregation over time, ending up with a sequence of spatial point patterns. Others require the underlying point process to be first-order separable, factorizing the estimates obtained for the spatial and temporal components of the intensity. In this regard, Mateu et al. (2020) introduces separable and pseudo-separable Voronoi-based intensity estimators; Moradi and Mateu (2020) presents a nonparametric and separable quick kernel-based intensity estimator.

We propose a novel methodology for Spatio-Temporal (Density and) intensity Estimation with Partial Differential Equation regularization (STDE-PDE) that permits to estimate the intensity of inhomogeneous Poisson point processes, evolving over time on a linear network spatial domain. We extend to the space-time setting here considered, the method developed by Clemente et al. (2023) for purely spatial density estimation on linear networks. Specifically, the inclusion of the temporal dimension follows the approach explored by Begu et al. (2024), who develop Space-Time Density Estimation with PDE regularization (see also Begu et al., 2023; Panzeri et al., 2023, and references therein), and leverages an equivalence between density estimation from an i.i.d. sample and intensity estimation in inhomogeneous Poisson point processes.

The proposed method penalizes the maximum likelihood intensity estimator with roughness penalties in space and time, such as those employed for instance in Bernardi et al. (2017) in the regression context. Due to the underlying mathematical model, STDE-PDE lies in the class of nonparametric PDE-penalized statistical models reviewed by Sangalli (2021) for spatial and functional data over complex domains. Unlike existing approaches, STDE-PDE does not aggregate data over time, nor does it assume any space-time separability of the point process. This significantly broadens the range of point patterns that the proposed technique can handle.

2 Model

Let $G := (W, E)$ be an undirected and compact graph embedded in a plane in $\mathbb{R}^2$. Here, $W := \{\mathbf{v}_k\}_{k=1}^K$ and $E := \{e_s\}_{s=1}^S$ denote respectively the sets of vertices and of edges, namely paired vertices, of G . Let X be a spatio-temporal inhomogeneous Poisson point process on the product space $G \times [0, T]$, with $T > 0$, and $N_X(A \times B)$ be the number of points of a set $(A \times B) \cap X$, for any $A \times B \subseteq G \times [0, T]$. Let $\gamma : G \times [0, T] \rightarrow \mathbb{R}^+$ be the inhomogeneous intensity function of the point process X , whose integral over $A \times B$ is the expected value of $N_X(A \times B)$. Our research goal is to estimate the unknown intensity from n i.i.d. realizations $\{(\mathbf{p}_i, t_i)\}_{i=1}^n \subseteq G \times [0, T]$ of X .

To this aim, we propose to compute the target spatio-temporal intensity estimate as the minimizer of the penalized negative log-likelihood of the considered inhomogeneous Poisson point process. We parameterize the estimation problem in terms of the log-intensity $\phi := \log \gamma$ to naturally comply with the non-negativity constraint on γ . The introduction of the regularization term, as the sum of roughness penalties in space and time, controls the regularity of the estimate, thereby preventing degenerate solutions associated with unbounded likelihoods. The contribution of such penalties is modulated by positive smoothing parameters, which balance the trade-off between the adherence of the model to the data and the smoothness of the estimate.

In order to theoretically validate the model, we first demonstrate the equivalence between intensity and density estimation problems. We extend the results of classical nonparametric literature, such as those achieved by Silverman (1982); Gu and Qiu (1993), to the case of interest. After defining the appropriate functional embedding, we formally establish the well-posedness of the considered estimation problem and derive the asymptotic properties of the proposed estimator.

Given the lack of a closed analytical form of the minimizer of the objective functional, we approximate it exploiting advanced techniques in the fields of optimization and numerical analysis. We develop a suitable discretization procedure that combines finite elements in space and B -splines in time. This ensures flexibility on undirected linear networks of any kind, enabling the handling of complex geometries. Figure 2 displays a spatial discretization of the road network of Bergamo considered in the application study.

To solve the minimization problem in its discretized form, we resort to optimized iterative routines, such as the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm. This contributes to maintaining feasible computational costs, even when dealing with a large dataset or a fine discretization of the spatio-temporal domain of interest. The contribution here presented further extends the utilities of the R/C++ `fdapde` (Arnone et al., 2023) for physics-informed spatial and functional data analysis, providing an effective tool for data evolving on linear networks over time.

3 Application to Road Accident Data

As an illustrative example, we present an application to real data concerning road accidents that occurred in Bergamo, Italy, from 2015 to 2022. Figure 1 shows some temporal snapshots of the observed point pattern. Figure 3 displays the corresponding intensity estimate obtained using STDE-PDE. This estimate exhibits spatially-dependent behaviors, characterized by high values along major arterial roads and in the busiest areas located in the city center. Additionally, the estimation evolves over time as expected. The proposed method proves capable of capturing spatio-temporal intensities of inhomogeneous Poisson point processes, successfully addressing all considered scenarios in this work.



Fig. 2. Spatial mesh of the road network of Bergamo, with 3 677 nodes and 4 525 edges.

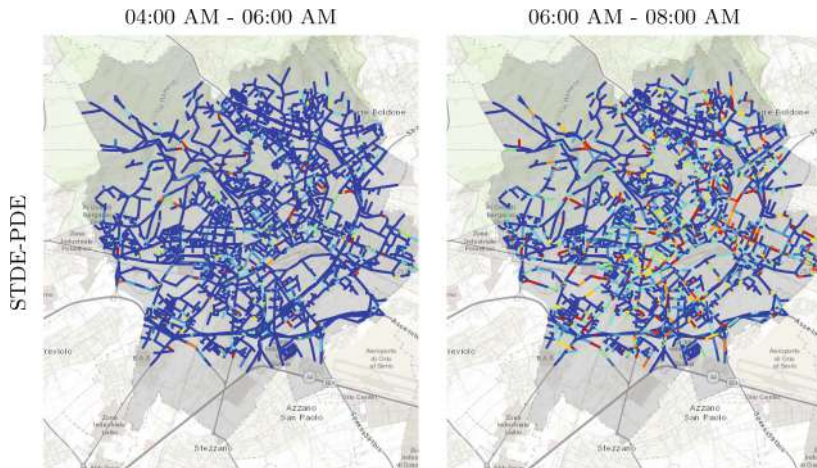


Fig. 3. STDE-PDE intensity estimate of road accidents occurred in Bergamo, Italy, from 2015 to 2022. The presented method returns accurate results, highlighting criticality in some areas of the city, especially during rush hours.

Acknowledgments. L.M. Sangalli acknowledges funding by the European Union - NextGenerationEU, in the framework of the projects GRINS - Growing Resilient, Inclusive and Sustainable (GRINS PE00000018 - CUP D43C22003110001), and PRIN2022 CoEnv - Complex Environmental Data and Modeling (CUP 2022E3RY23). We also acknowledge the MUR research project Dipartimento di Eccellenza 2023-2027, Dipartimento di Matematica, Politenico di Milano.

References

- Arnone, E., Clemente, A., Sangalli, L.M., Lila, E., Ramsay, J., Formaggia, L.: fdaPDE: Physics-informed spatial and functional data analysis (2023). <https://CRAN.R-project.org/package=fdaPDE>. (R package version 1.1-17)
- Begu, B., Panzeri, S., Arnone, E., Carey, M., Sangalli, L.M.: A nonparametric penalized likelihood approach to density estimation of space-time point patterns. *Spatial Stat.* (2024). <https://doi.org/10.1016/j.spasta.2024.100824>
- Begu, B., Panzeri, S., Arnone, E., Sangalli, L.M.: A novel spatio-temporal estimation method for occurrences over planar and curved regions, in: *Proceedings of the GRASPA 2023 Conference*, pp. 107–112 (2023). ISBN: 979-12-210-3389-2
- Bernardi, M.S., Sangalli, L.M., Mazza, G., Ramsay, J.O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic Environ. Res. Risk Assessment* **31**, 23–38 (2017). <https://doi.org/10.1007/s00477-016-1237-3>
- Clemente, A., Arnone, E., Mateu, J., Sangalli, L.M., et al.: Non-parametric density estimation over linear networks. In: *Proceedings of the GRASPA 2023 Conference*, pp. 141–145 (2023). ISBN: 979-12-210-3389-2
- Diggle, P.J.: *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press (2013). ISBN: 978-10-324-7747-3
- González, J.A., Rodríguez-Cortés, F.J., Cronie, O., Mateu, J.: Spatio-temporal point process statistics: a review. *Spatial Stat.* **18**, 505–544 (2016). <https://doi.org/10.1016/j.spasta.2016.10.002>
- Gu, C., Qiu, C.: Smoothing spline density estimation: theory. *Annals Stat.* **21**, 217–234 (1993). <https://doi.org/10.1214/aos/1176349023>
- Mateu, J., Moradi, M., Cronie, O.: Spatio-temporal point patterns on linear networks: Pseudo-separable intensity estimation. *Spatial Statist.* **37** (2020). <https://doi.org/10.1016/j.spasta.2019.100400>
- Moradi, M.M., Mateu, J.: First-and second-order characteristics of spatio-temporal point processes on linear networks. *J. Comput. Graph. Stat.* **29**, 432–443 (2020). <https://doi.org/10.1080/10618600.2019.1694524>
- Okabe, A., Yamada, I.: The K-function method on a network and its computational implementation. *Geograph. Anal.* **33**, 271–290 (2001). <https://doi.org/10.1111/j.1538-4632.2001.tb00448.x>
- Panzeri, S., Begu, B., Arnone, E., Sangalli, L.M.: An estimation tool for spatio-temporal events over curved surfaces. In: *Proceedings of the Statistics and Data Science Conference*, pp. 201–206 (2023). ISBN: 978-88-6952-170-6
- Sangalli, L.M.: Spatial regression with partial differential equation regularisation. *Inter. Stat. Rev.* **89**, 505–531 (2021). <https://doi.org/10.1111/insr.12444>
- Silverman, B.W.: On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Stat.* **10**, 795–810 (1982)



Press Coverage of the Economy: Construction of Coverage and Sentiment Indices

Paolomaria Fabrizio^(✉), Patrizio Lodetti, Raffaele Guetto, and Daniele Vignoli

Department of Statistics, Computer Science, Applications “G. Parenti”,
University of Florence, Viale Morgagni 59, 50134 Firenze, Italy
{paolomaria.fabrizio, patrizio.lodetti, raffaele.guetto,
daniele.vignoli}@unifi.it

Abstract. Media represents the main source of information regarding economic issues. This article builds comprehensive indices reflecting the media narrative on the overall state of the economy in Italian newspapers using the Lexis Nexis database from 2009 to 2020. For this reason, we employ a machine learning approach, implementing two NLP models with the language model BERT: the first discriminates between articles discussing the economy and those that do not, while the second assigns sentiment to the “in-topic” news. Additionally, we compare these coverage and sentiment indices with traditional indicators of the Italian economy. The results show that the economic trends described in the media are not always aligned with actual economic trends. We also discuss potential applications of these indices in population studies, such as fertility research, highlighting the potential of media-based indices as a complementary tool for understanding social phenomena.

Keywords: Uncertainty · Bert · Sentiment Analysis · Indices

1 Introduction and Aim

Since the financial crisis between 2008 and 2010, the global economy has experienced a recovery and a subsequent expansion phase till 2020, where economic tension between the USA and China, Brexit and Covid 19 pandemic have been the causes of a new and economic sharp slowdown. These events have raised concerns about uncertainty of the future, contributing to this slowdown (Baker et al., 2016). The economy and the economic subjects are driven by expectations about the future and by rational decisions involving a cost-benefit approach, which can be undermined by a context of uncertainty. As discussed by Vignoli et al. (2020) economic uncertainty should be estimated taking heed of that individuals produce their own narrative of the future considering structural constraints, their own expectations and imaginaries. These narratives are the result of the socio-cultural environment that surrounds them and the media opinion formers

with which they get in contact. Understanding economic uncertainty is troublesome because it is not directly observable through an interpretation of the classic economic indicators. In recent years, there has been a growing interest in constructing indices based on media narratives to supplement traditional economic indicators. These indices aim to capture the overall sentiment and themes expressed in the media coverage of the economy. The reason behind the construction of these indexes is due to the drawbacks related to the traditional measures that could not reflect how individuals perceive socioeconomic phenomena. Press coverage-based and sentiment-based indicators, on the contrary, allow to explore the media salience of economic narratives which may have a greater impact on behaviours than traditional indices. The aim of this study is to create comprehensive indices reflecting the media narrative of the economy exploiting a machine learning approach. This is done through the implementation of two Natural Language Processing models with the language model BERT (Devlin et al., 2019): the first model discriminates between in-topic (talking about economy) and off-topic news, whereas the second model assigns a sentiment (positive, neutral, or negative) to the in-topic news. In addition to the construction of these indices, this article also examines the reliability and validity of the constructed indices and discuss their potential applications. Despite the existence of literature on constructing indices that measure media coverage and sentiment, there is a noticeable gap when it comes to developing such indices for economic news using machine learning approaches like BERT. While previous studies have successfully employed machine learning models to analyze sentiment and coverage in various domains, including finance and politics, the specific domain of economic news in Italy remains relatively unexplored in terms of applying advanced natural language processing techniques.

2 Data

The indices were constructed using data sourced from Lexis Nexis that is a unique database containing a comprehensive collection of news articles from various newspapers from different European countries. The data provided by Lexis Nexis offers a diverse array of news originating from multiple sources and media types, including national newspapers, local newspapers, online newspapers, and news agencies. We extracted 18358800 articles belonging to Italian newspapers and news agencies that range from January 2009 to June 2020. The absolute number of articles is not constant over the period of analysis. The greatest journalist coverage occurs in the years 2010, 2011 and 2012, while there is a lower number of articles in 2020 because the observation period is until June.

3 Analysis and Methodological Approach

To construct the aforementioned indices, first we developed guidelines for manually annotating the data. Next, we extracted a sample of almost 2000 articles from the Lexis Nexis database, which were then annotated by experts according

to these guidelines. Finally, the annotated sample was used to fine-tune a pre-trained machine learning model that uses the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), to predict annotations on the full dataset. The pre-training model used is the *dbmdz/bert-base-Italian-xsl-cased* that has been trained by the MDZ Digital Library team (dbmdz) at the Bavarian State Library and made it available on the Hugging Face platform.

The guidelines to interpret news included rules for distinguishing between economic and non-economic articles, and for classifying economic articles as negative, positive, or neutral. Specifically, an economic *article* is considered as such if the prevalent topic regards economic issues like unemployment, employment, salary, crisis, prices, etc. A negative sentiment has been assigned to an economic article if it reports *negative* facts or could fuel *negative* perceptions about the economy/economic conditions. A positive sentiment has been assigned to an economic article if it reports *positive* facts or could fuel *positive* perceptions about the economy/economic conditions. And a neutral sentiment has been assigned to an economic article if it has neither negative nor positive or if it is unclear (has both positive and negative sentiment).

3.1 The Economic Coverage Index

The first index we constructed, the **Economic Coverage Index** (ECI), follows the Economic Policy Uncertainty index construction of Baker et al. (2016): first we count, for each month of analysis, the proportion of articles in a given month that the BERT model classifies as economic. These raw counts are not a precise measure and should be contextualized with respect to the newspaper, time period, and other news written by the same newspaper during that time. For this purpose we divided by the total number of articles in the same newspaper and month: ($\frac{\text{\#articles talking about economy}}{\text{\#total articles}}$); then we standardized each newspaper time series to unit standard deviation and averaged.

To train the model we used 1990 articles, 1209 did not refer to economy and 781 were related to economic themes.

The training set was split into a 75/25 ratio for training and testing purposes. Afterwards, the test set was further divided into a 50/50 ratio to create a validation set and a final test set. Finally the training process returned an f1-score of 90% as shown in Table 1.

Table 1. In-OffTopic model performances

	precision	recall	f1-score
OffTopic	0.91	0.92	0.92
InTopic	0.87	0.86	0.86
Total			0.90

3.2 The Economic Sentiment Index

The second index, the **Economic Sentiment Index** (ESI), is built according to the same approach but, in this case, we count, for each month of analysis, the articles in a given month that the BERT model classifies as positive minus the ones classified as negative scaled by the sum of the positive plus the negative ones $\frac{(\#pos - \#neg)}{(\#pos + \#neg)}$, excluding the neutral ones. Negative and positive sentiment are both considered in a unique index but neutral news is excluded to simplify the index calculation and more effectively highlight variation between positive and negative sentiments (Guetto et al. (2023); Soo (2018)). The training phase involved 2927 articles, 623 negative, 623 neutral and 1680 positive. The performances of the model are reported below in Table 2.

Table 2. Sentiment model performances

	precision	recall	f1-score
Negative	0.63	0.63	0.63
Neutral	0.43	0.45	0.44
Positive	0.8	0.79	0.8
Total			0.69

In Fig. 1 a representation of the ECI and ESI indices for Italy.

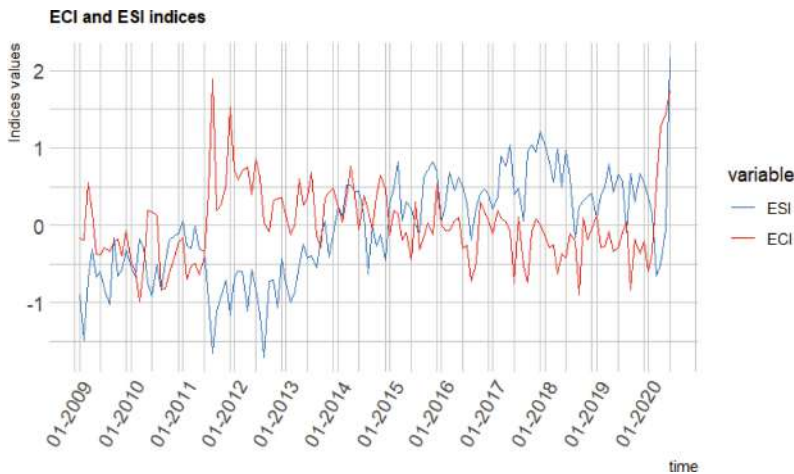


Fig. 1. Economic Coverage Index and Economic Sentiment Index

The ECI has its high points in 2011 and 2020 i.e. these are the times when newspapers attention given to the economy increased. Instead, it might be interesting to study the existence of some association between ECI and some real

economy indices (Table 3). The ESI shows an inverse relationship with the ECI, with a correlation coefficient of -0.19. Specifically, there are certain periods, such as 2011 and the pre-pandemic era, where the ECI registers high values while the ESI remains low. Conversely, from 2015 to 2019, the trend reverses. It seems that when the volume of economic news increases, the tone of the news is not optimistic. The ECI index seems to capture some economic aspects of the GDP. The ESI, opposite to ECI, is associated more with subjective economic indices such as the consumer confidence climate, the present consumer climate and the future consumer climate. However, we have not found an index that closely tracks our indices or that is highly correlated with them, which means that our indices captures something different from other traditional economic indices.

Table 3. Eci and Esi correlations with traditional economic indices

	GDP	Un rate	Ind Prod	CCC	PCC	FCC
ECI	-0.61	0.19	0.24	-0.41	-0.46	-0.39
ESI	-0.0001	0.36	0.015	0.65	0.62	0.75

4 Conclusion

In this study, we show how to construct two press-based indices, the Economic Coverage Index (ECI) and the Economic Sentiment Index (ESI), to gauge the dynamics of economic news coverage and sentiment in Italy. The ECI, derived from our classification model and constructed following the methodology outlined by Baker et al. (2016), reveals heightened press attention to the economy during 2011 and 2020. We did not find a strong correlation between this index and real economy index (despite on GDP variations), suggesting that the ECI captures unique facets of economic discourse and that the media are not always aligned with actual economic trends. The ESI, constructed similarly but focusing on sentiment analysis (using the BERT model), displays an inverse relationship with the ECI. Notably, during periods of high economic news coverage, the sentiment conveyed tends to be less optimistic. The ESI correlates more strongly with subjective economic indices, suggesting its association with perceptions rather than objective economic indicators. These considerations emphasise the importance of developing indicators that capture the perception of the economy, which is fundamental to the creation or deconstruction of an imaginary of uncertainty with the consequent impact on people’s attitudes and behaviour.

References

Baker, S.R., Bloom, N., Davis, S.J.: Measuring Economic Policy Uncertainty*. Quarterly J. Econ. **131**(4), 1593–1636 (2016). <https://doi.org/10.1093/qje/qjw024>, <https://academic.oup.com/qje/article-pdf/131/4/1593/30636768/qjw024.pdf>

- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. 1810.04805 (2019)
- Guetto, R., Morabito, M.F., Vollbracht, M., Vignoli, D.: Fertility and media narratives of the economy: evidence from italian news coverage. *Demography* **60**(2), 607–630 (2023). <https://doi.org/10.1215/00703370-10607928>, <https://read.dukeupress.edu/demography/article-pdf/60/2/607/1842452/607guetto.pdf>
- Soo, C.K.: Quantifying Sentiment with News Media across Local Housing Markets. *Rev. Financial Stud.* **31**(10), 3689–3719 (2018). <https://doi.org/10.1093/rfs/hhy036>, <https://academic.oup.com/rfs/article-pdf/31/10/3689/25693090/hhy036.pdf>
- Vignoli, D., Guetto, R., Bazzani, G., Pirani, E., Minello A.: A reflection on economic uncertainty and fertility in europe: the narrative framework. *Genus* 76 (2020). <https://doi.org/10.1186/s41118-020-00094-3>



“Stay Tuned!”: Validating a Questionnaire on Perceptions of University Students in Staying Informed with Current Events

Anna Maria Parroco , Giulia Bagnasco, Morena Cucinella, Santina Pampillonia, and Micaela Arcaio  

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

{annamaria.parroco,micaela.arcaio}@unipa.it

Abstract. Information behaviour has been widely discussed in the literature, given the interest in how and why human beings seek and use information coming from librarians, policy-makers and social scientists. Understanding information consumers’ needs also provides the chance for practical applications to computer scientists and cognitive psychologists. In this study, information behaviour – related to current political news and events – is shaped within the framework of the theory of reasoned action, highlighting whether behavioural beliefs and normative beliefs influence behavioural intentions, which influence behaviour itself. The aim is, thus, to evaluate the internal consistency and factorial validity of an ad-hoc questionnaire administered to university students on information behaviour. The internal consistency is checked using the Cronbach’s alphas of the three dimensions involved, while the factorial validity is assessed using a structural equation model. The results consist of a short questionnaire that could be employed to understand the information behaviour of university students to offer experts of information creators and disseminators an instrument to study their public’s needs.

Keywords: Information Behaviour · Theory of Reasoned Action · Factorial Validity · Current events · News

1 Information Behaviour

Information behaviour has been widely discussed in the literature. The interest in how human beings interact with, seek and use information comes from several sources: from librarians trying to make sense of how the public perceives their services, to policy-makers seeking to learn how new scientific findings make their way into the economy, to social scientists who are generally interested in the way information is used as a social tool [1]. So, in the current information society, the careful design of information systems and services aligned with consumers’ needs could offer prospects to develop predictive models and practical applications in the fields of computer science and cognitive psychology.

Ajzen and Fishbein’s model [2] has been proven effective in predicting people’s behavioural intentions in various situations [3, 4], but it remains untested to evaluate the

need for information. This model is based on behavioural and normative beliefs, which influence intentions and then people’s actual behaviour.

Numerous studies support the hypothesis that behavioural beliefs, i.e., expectations and values attributed to one’s behaviour, influence the need for information [5]. Similarly, it is highlighted that behavioural intention regarding information influences people’s actual behaviour in terms of searching for and using information [6].

Building on this premise, a research activity was initiated to evaluate, for the specific target group of young university students, whether the relationships between beliefs, norms, and behavioural intentions explain information behaviour related to current political news and events. An ad-hoc questionnaire was, thus, constructed to assess information seeking behaviour among young university students. As far as we know, a specific instrument based on the theory of reasoned action has not yet been proposed in the literature for this socio-demographic target.

In this study, the focus is on the validity of the scales used in the proposed questionnaire, which represent the dimensions involved in information behaviour.

The structure of the questionnaire will be presented in Sect. 2, along with the methodology employed to evaluate the internal consistency of all the identified dimensions; in Sect. 3, the results of the validation process will be presented; Sect. 4 will contain some remarks, as well as limitations and further developments of this work.

2 The Questionnaire and Its Validation Procedure

2.1 The Information Behaviour Questionnaire

Structure The questionnaire to be used for the “Information Behavior Survey” (IBS) is divided into four sections exploring expectations, attitudes, values, and behaviours related to information consumption. All the items have been constructed using a 7-point Likert scale, going from “Completely disagree” to “Completely agree”. The four sections match the dimensions identified by the theory of reasoned action [2]. Here, only the dimensions involved in the factorial validity are discussed. These are:

- Behavioural beliefs, linking the desired behaviour to outcome expectations, and emphasizing the subjective likelihood that the behaviour itself leads to specific experiences or outcomes. These beliefs, coupled with the values attributed to expected outcomes, influence the overall attitude toward behaviour. This dimension is made of 13 items, e.g., “Keeping oneself up to date with news helps express one’s opinions and values”.
- Normative beliefs, encompassing the perception of pressure coming from prominent individuals or groups (such as spouses, family members, friends, and other social references) to adopt specific behaviours. Five items make up this dimension, e.g., “I believe that my family thinks it’s important to keep up with the news”
- Behavioural Intentions, about whether people are willing to perform a specific behaviour – and it is the direct precursor of behaviour itself. Four items contribute to this dimension, e.g., “I mean to keep up with the news”.

Sample, Sample Size and Other Considerations. The sample design of the IBS survey is not random, as the usual definition would make it. Respondents were, indeed, recruited during one of their classes: thus, the only source of randomness is given by whether they were present that day or not.

Cronbach's alpha was used to determine the sample size for the pilot survey: for a generic dimension of 15 items, with a 5% significance, power set to 80.0%, and a null hypothesis of Cronbach's alpha = 0.5 vs. an alternative hypothesis of Cronbach's alpha = 0.7, the sample size was estimated at 67 subjects minimum [7]. Seventy-one students were thus administered the pilot survey while 279 students were recruited for the main survey.

Both the pilot and the main survey were administered only after careful consideration of an ethical committee: privacy policies were enforced to grant the participants' rights of anonymity as well as to avoid any source of risk or discomfort.

Administration. The questionnaire was administered during scheduled academic hours, on different days, using an online survey tool, which the respondents used on their own electronic devices. Its final version – whose validation is presented here – is the result of the pilot survey conducted on a smaller sample (of 71 students). Feedback from this smaller sample allowed to clean up the first version of the items, as well as checking for possible sources of unusual formatting and difficulty in completing the questionnaire.

2.2 The Factorial Validity

Methods. The validation process is made of several steps. First, the internal consistency of the dimensions – as discussed before – is checked then the factorial validity of the model is provided.

Cronbach's alpha is used to evaluate the internal consistency of the dimensions in the theoretical model. Levels of alpha higher than 0.4 make the items that contribute to a specific dimension “quite reliable”, while values of alpha higher than 0.6, make the dimension “reliable” [8].

A structural equation model (SEM) is, then, estimated to assess the factorial validity of the questionnaire [9]. “Actual behaviour” was measured on a seven-point Likert scale as a manifest endogenous variable about self-perception of information behaviour (“I believe I keep up with the news”). The model's structural component can be found in Fig. 1.

This model was estimated twice: in the first round, the model's goodness of fit already satisfied the common criteria of evaluation; but, given the presence of items whose standardised loading was lower than the usual 0.4 threshold, the model was estimated once more without these items.

Since all items are Likert scales, the DWLS estimator was used [10]. SEM's usual goodness-of-fit statistics are used – the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardised Root Mean Square Residual (SRMR). R 4.3.2 [11] and the lavaan [12] package were employed for the estimation of the SEMs.

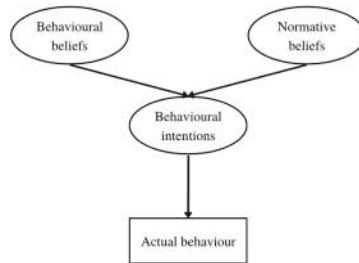


Fig. 1. The model’s structural component.

3 Results

The validation process was conducted on sample made of 279 university students (both from Bachelor’s and Master’s degree courses), with an average age of 22.32 years ($SD = 4.72$). The sample is predominantly composed of girls – 79.2% vs 18.3% of boys, while the rest either prefer not to respond or declare “other”.

According to the results of the Cronbach’s alphas (see Table 1), all the constructs involved in the questionnaire can be considered either “quite reliable” or “reliable”. The dimension with the lowest reliability is “normative beliefs”, while “behavioural beliefs” and “behavioural intentions” roughly scored the same.

Table 1. Cronbach’s alphas for the constructs in the questionnaire.

Construct	Cronbach’s alpha
Behavioural beliefs	0.64
Normative beliefs	0.55
Behavioural intentions	0.65

As specified above, with the first step, the items whose standardized factor loading was lower than the 0.4 threshold have been excluded from the model. This reduced the number of items both in the dimension of “behavioural beliefs”, from thirteen to six, and the number of items in “normative beliefs”, from five to three. All items in “behavioural intentions” had standardized loading higher than 0.6 – thus, this dimension was not modified. This process did not significantly change the Cronbach’s alphas as listed above.

The new round of estimation for the SEM allowed to check for the relationships between these constructs. The evaluation of the goodness-of-fit statistics suggests an overall good fit of the model ($CFI = 0.937$; $RMSEA = 0.092$; $SRMR = 0.087$), whose values point to a better fit. However, not all the relationships between the constructs as shown in the theoretical structural model shown in Fig. 1 were significant: while “behavioural beliefs” seem to be quite significant in determining “behavioural intentions” (St. Coefficient = 0.924; $p\text{-value} < 0.001$) – which, in turn, resulted as significantly

associated with the actual behaviour (St. Coefficient = 0.495; p -value < 0.001) - “normative beliefs” are not related to “behavioural intentions” (St. Coefficient = -0.024; p -value = 0.894). These results point to the idea that their own opinions and their own beliefs are what matters to the students in the sample – and that what other people think has no relevance in whether they keep up with the news or not.

4 Discussion and Conclusion

The goal of this study is to assess the validity of a questionnaire on information behaviour, whose dimensions are based on the theory of reasoned action [2]. According to this theoretical framework, information behaviour should be explained by the intention to do so (“behavioural intentions”), as well as by the value and expectations attributed by oneself to information (“behavioural beliefs”) and by peer pressure to keep up with the news (“normative beliefs”).

The validation procedure employed here shows that the three scales used to represent the dimensions in the framework have either quite reliable or reliable internal consistency, and the SEM estimated for the factorial validity presented statistics pointing to an overall good fit of the model, thus demonstrating that the questionnaire could be considered a satisfactory instrument to evaluate information behaviour in young university students.

While beyond the scope of this preliminary analysis, the results of the SEM can also be used to explore whether “behavioural beliefs” and “normative beliefs” are associated with “behavioural intentions”, and how this last construct is associated with the actual (self-perceived) behaviour of respondents. The only non-significant relationship is that between “normative beliefs” and “behavioural intentions”. This result suggests that, for the young participants in the study, personal opinions and beliefs hold more weight than external expectations, emphasising the autonomous and self-driven nature of the information-seeking behaviour. This result is confirmed by an extensive review of information behaviour that states that people don’t “need” information, but “want” information, thus showing the relevance of personal motivation in seeking information [5].

These results lead to consider this ad-hoc questionnaire as an apt tool to study information behaviour among university students. The validation process carried out here highlights which items and which dimensions researchers of information behaviour should focus on. A relevant advantage of this instrument lies in its simplicity, requiring minimal time for completion and a reduced attention span: this, alongside its easy structure, allows it to be self-administrable.

This work does not, of course, lack limitations. First and foremost, the questionnaire was validated only on university students. Broader demographics, starting with young adults not enrolled in a degree course, may require additional adjustments and considerations. Another possible limit may be the timing of the administration of the questionnaire itself: in this case, the data were collected in the final two months of 2023, during which different Italian and global events and news might have urged students to seek information more frequently and attentively.

Finally, many further developments are already underway, including a more in-depth evaluation of the relationships between the different dimensions involved in information

behaviour – so a more comprehensive understanding of this phenomenon can be provided to those interested in the mechanisms that push people into seeking information.

References

1. Bates, M.J.: *Encyclopedya of Library and Information Sciences*, Fourth Edition, Taylor and Francis (2010)
2. Fishbein, M., Ajzen, I.: *Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research*. Addison-Wesley, MA (1975)
3. Richardson, B., Wang, Z., Hall, C.: Blowing the whistle against greek hazing: the theory of reasoned action as a framework for reporting intentions. *Commun. Stud.* **63**(2), 172–193 (2012)
4. Fitzmaurice, J.: Incorporating consumers’ motivations into the theory of reasoned action. *Psychol. Mark.* **22**(11), 911–929 (2005)
5. Case, D.O., Given, L.M.: *Looking for Information: A survey of research on information seeking, needs, and behavior* (2016)
6. Browne, G.J., Cheung, C.M.K., Heinzl, A., Riedl, R.: Human Information Behavior. *Bus. Inf. Syst. Eng.* **59**(1), 1–2 (2016). <https://doi.org/10.1007/s12599-016-0458-9>
7. Bujang, M.A., Omar, E.D., Baharum, N.A.: A review on sample size determination for Cronbach’s alpha test: a simple guide for researchers. *Malays. J. Med. Sci.* **25**(6), 85–99 (2018)
8. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: *Multivariate Data Analysis*, 7th edn. Pearson, New York (2010)
9. Muthén, B.: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**(1), 115–132 (1984)
10. Muthén, B., du Toit, S.H.C., Spisic, D.: Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished technical report (1997)
11. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. Last accessed 4 Mar 2024
12. Rosseel, Y.: Lavaan: an r package for structural equation modeling. *J. Stat. Softw.* **48**(2), 1–36 (2012)



Ordinal Discriminative Dimensionality Reduction of Functional Profiles

Giulia Patanè¹(✉), Federica Nicolussi¹, Alexander Krauth³, Günter Gauglitz², Bianca Maria Colosimo², Luca Dede¹, and Alessandra Menafoglio¹

¹ MOX, Department of Mathematics, Politecnico di Milano, Milan, MI, Italy
{giulia.patane,federica.nicolussi,luca.dede,
alessandra.menafoglio}@polimi.it

² Department of Mechanics, Politecnico di Milano, Milan, MI, Italy
guenter.gauglitz@uni-tuebingen.de, biancamaria.colosimo@polimi.it

³ University of Tübingen, Tübingen, Tü, Germany
alexander.krauth@uni-tuebingen.de

Abstract. Optical biosensors, utilizing biological components like DNA and antibodies, are effective analytical tools. However, analyzing sensor signals, especially in large datasets such as reflectometric imaging sensors, is computationally demanding both in terms of time and of memory usage. In this communication, we employ data from a reflectometric imaging sensor to track the antibody-antigen reaction progression using video images of the biosensor surface. Analyzing temporal changes in light intensity provides insights into reaction progression; however, the need for an automatic detector of biological process reaction requires dimensionality reduction of the data from the sensor, which arrive in the form of functional profiles from video signals. We illustrate a workflow which includes cleaning light disturbances, condensing video data, and reducing the dimensionality of the obtained functional data. Departing from traditional methods like Functional Principal Component Analysis (FPCA), we discuss functional-ordinal Canonical Correlation Analysis to optimize correlation between the high-dimensional predictor and the outcome. The conclusion highlights that, the combination of preprocessing and CCA for effective discrimination among different reagent concentration levels, enables one to project the video signal into a 2-dimensional space. This innovative approach enhances our ability to detect virus vitality in biomanufacturing processes.

Keywords: Functional Data Analysis · Ordinal Data · Canonical Correlation Analysis · Biosensors

1 Introduction

Optical biosensors are a powerful analytical tool for diagnostic purposes. As a sensing element biological components such as DNA, antibodies, enzymes, cell receptors, and even whole cells can be used [14–16]. In the case of large data

sets - for example in reflectometric imaging sensors - the analysis of the sensor signals can be very time-consuming and requires huge amounts of memory space while the output of these analyses often is a single parameter, for example, the concentration of the analyte. In the present study, a reflectometric imaging sensor was used to monitor the progression of the reaction between an antibody and an antigen through video images of the biosensor surface. Here, the statistical challenge consists in analysing the temporal profiles of the light intensity (i.e., functional observations) in Regions of Interest (ROIs), in response to the binding of the antibody to the antigen spot of pre-specified concentrations onto the sensor surface. Given the necessity for simultaneous and real-time identification of reactions across numerous ROIs, an automatic, efficient and fast reaction detector becomes imperative. This underscores the need for dimensionality reduction in sensor data, typically presented as video signals. In this communication, we introduce the dimensionality reduction method proposed in [12], with a specific emphasis on discerning the presence of a reaction in a designated region and on predicting the reagent concentration level. For this purpose, we illustrate a pipeline composed of three steps. Since our focus is on video signals, the initial phase involves cleaning of the light disturbances. The subsequent phase is dedicated to extracting a functional profile from the video frames corresponding to the ROI. The final phase centers on reducing dimensionality of the functional data. Regarding the last step, traditional literature suggests employing Principal Component Analysis (PCA) for dimensionality reduction in predictive models, where the first k components are utilized as predictors. While PCA seeks components maximising the variance, we here follow [12] and conduct the dimensionality reduction by optimizing the correlation between the high-dimensional predictor and the outcome. In this sense, we shall rely on a variant of Canonical Correlation Analysis named functional-ordinal CCA, to select components that effectively discriminate among units with different concentration levels.

2 Dataset and Preprocessing

The dataset consists of a video signal, recording one image per second, captured through a reflectometric sensor. Each video frame represents the light intensity of the reflected light. Specifically, 1035 spots, each containing a certain concentration of the antigen, are arranged in a regular grid on the sensor surface, identifying 1035 ROIs. In practice, some spots may not exist or contain minimal reagent concentration. Note that visually detecting the presence of a spot is straightforward, as it appears darker than the background in the reflectometric image. In this section we report the preprocessing of the video signal, from the light cleaning to the construction of the final dataset, where the statistical unit is the ROI and the covariates are (i) a univariate functional datum that represents the light intensity temporal evolution and (ii) an ordinal variable that represents the level of concentration of reagent.

For each ROI, the nominal concentration level is associated with an ordinal label. In order to detect magnitude outliers among the ROIs, we eliminate exogenous sources of light from the reflectometric images, such as external lights and

irregularities in the surface coating. In our case study, it is realistic to assume that the exogenous sources of light are constant over time. We model the logarithm of light intensity at time 0 as a Generalized Additive Model (GAM) [6], using as predictor the pixels's spatial coordinates (x, y) , with a nonlinear contribution, and the categorical information of whether or not the pixel belongs to a ROI. By subtracting the model-based expected light disturbance on the reflectometric images, we effectively minimize variations in light that do not stem from the presence of the antigen spot. Net of the light cleaning, we can easily check if a pixel belongs to random spots or determine if a randomly sampled pixel is located within or outside a ROI.

Light intensity is a variable ranging from 0 to $2^{16} - 1$, making it compositional in its nature. To conduct standard statistical analyses (such as dimensionality reduction, models or clustering) it is essential to work with absolute information rather than relative, as is the case with compositional data [8]. Several transformations for compositional data are suggested in the literature to ensure meaningful and interpretable results; amongst the most popular methodologies, we mention that based on the log-ratio approach (firstly proposed by Aitchison, see, e.g., [4]) which is also used in [12]. Following these works, the univariate functional datum representing the light intensity of the i -th ROI is built as follows. We subtract from the raw signal the GAM model of light disturb, obtaining a cleaned time-varying light intensity, $f_j(t)$ for each pixel j in the image D . For each ROI i , we compute the 3%-trimmed mean of the (cleaned) signals $f_j(t)$ corresponding to $N = 100$ randomly extracted pixels within the ROI, obtaining $\bar{f}_i^{ROI}(t)$. Note that this guarantees robustness for the estimated mean function within the ROI, which is then used as representative of the signal for the ROI. We also sample $N = 100$ pixels in the neighborhood of each ROI i and we compute the 3%-trimmed mean $\bar{f}_i^{BG}(t)$, in order to have a robust estimation of the mean in the background area adjacent to the ROI i . In order to take into account for the compositional nature of the signals, we compute the log-ratio:

$$\bar{x}_i(t) = \log\left(\frac{\bar{f}_i^{ROI}(t)\bar{f}_i^{BG}(0)}{\bar{f}_i^{ROI}(0)\bar{f}_i^{BG}(t)}\right)$$

We smooth the log-ratio through smoothing splines regression [9] with 20 cubic splines, equally distributed knots, and penalty parameter equal to 10^5 selected via GCV, in order to obtain a meaningful functional profile. In conclusion, every ROI i is associated with a functional datum $x_i(t)$ and a level L_i of concentration of reagent. In Fig. 1 one can observe the functional boxplot of the log-ratio by level of concentration.

3 Dimesionality Reduction

We aim at reducing the dimensionality of the functional data while maximizing the ability to predict the ordinal level in the reduced space. Following [12], we rely on Canonical Correlation Analysis [7] which finds the directions that maximize the correlation between two datasets. In [3] the concept of CCA was extended

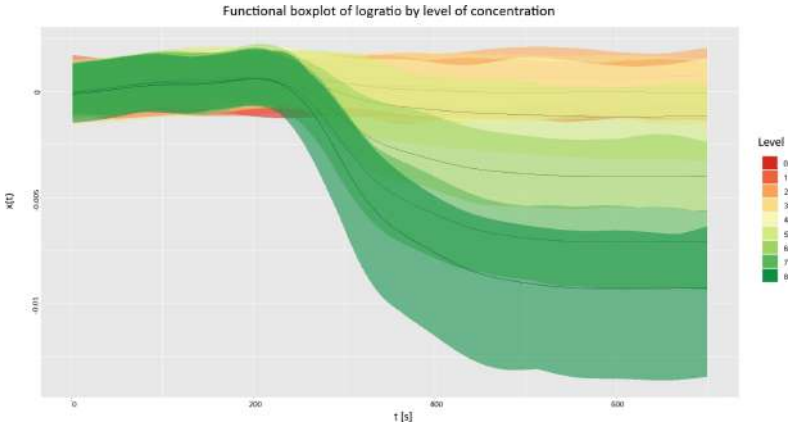


Fig. 1. Functional boxplot of a sample of smoothed log-ratio, stratified by concentration level.

to functional datasets [1], where the smoothness of the functional directions was imposed by introducing appropriate penalizations in the objective functional. As a further challenge, we here aim to use the concept of canonical correlation in a context in which one wants to find the functional canonical directions that enables one to discriminate an ordinal variable (the concentration), from the reduced functional signal (the time-varying preprocessed light intensity). By ordinal variable we here refer to discrete but not categorical variable for which labels have a well defined order [5]. For an introduction to a useful approach for their statistical analysis we refer to [13]; we here limit to recall that this case needs a suitable encoding of the ordinal variable, in order to take account the order and the possible non-equidistance of consecutive levels. In [12] we propose an encoding criterion for the ordinal variable, suitable to use CCA in the functional-ordinal framework, and we provide an interpretation of the canonical directions. On this basis, [12] develops a novel methodology, named functional-ordinal CCA, which allows one to identify the (smoothed) directions of maximum association between a functional and an ordinal dataset, robustly to the presence of magnitude and shape outliers [10][11]. In this communication, we shall discuss the functional-ordinal CCA methodology and its application for biosensors. In particular, the analysis of the available data suggests that a functional-ordinal CCA dimensionality reduction of dimension 2 allows to effectively discriminate among different levels of concentrations, unlike classical dimensionality reduction methods such as PCA.

4 Conclusions and Further Analyses

In this communication, we discuss the combination of the preprocessing showed in Sect. 2 with dimensionality reduction through the functional-ordinal CCA, in

which we encode the ordinal datum in a suitable way that allows the levels to be non-equidistant. This strategy allows one to project a high dimensional object such as the video of light intensity into a lower dimensional space, maximizing the ability to discriminate among different levels of the ordinal variable. In [12] we validate the functional-ordinal CCA methodology, through k-fold Cross Validation. Moreover, we show through extensive simulations that functional-ordinal CCA out-performs the FPCA, in terms of classification accuracy, even in situations for which the dataset shows a higher variability than that explained by the ordinal variable only, as it is typically the case for the class of target datasets in the biosensor applications considered in this work.

Acknowledgements. This research has received funding by the European Commission under the “HORIZON-CL4-2021-DIGITALEMERGING-01 project BioProS - Biointelligent Production Sensor to Measure Viral Activity” (grant agreement no. 101070120), 2022-2026. The present research is part of the activities of “Dipartimento di Eccellenza 2023-2027”, MUR, Italy, Dipartimento di Matematica, Politecnico di Milano.

References

1. Leurgans, S.E., Moyeed, R.A., Silverman, B.W.: Canonical correlation analysis when the data are curves. *J. Roy. Stat. Soc.* **55**, 725–740 (1993)
2. Silverman, B.W.: Smoothed functional principal components analysis by choice of norm. *Ann. Stat.* **24**, 1–24 (1996)
3. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. Springer (2005)
4. Aitchison, J.: *The Statistical Analysis of Compositional Data*. Monographs on statistics and applied probability. Springer (2011)
5. Agresti A.: *Categorical Data Analysis* (3 ed.). John Wiley & Sons, Hoboken (2013)
6. Hastie, T., Tibshirani, R.: Generalized additive models. *Statist. Sci.* **1**(3), 297–310 (1986)
7. Härdle, W., Simar, L.: *Canonical Correlation Analysis*. Applied Multivariate Statistical Analysis. Springer (2007)
8. Greenacre M.: Compositional data analysis. *Annual Rev. Stat. Appli.* (2021)
9. Reinsch, C.H.: Smoothing by spline functions. *Numer. Math.* **10**, 177–183 (1967)
10. Lopez-Pintado, S., Romo, J.: A half-region depth for functional data. *Comput. Stat. Data Anal.* **55**, 1679–1695 (2012)
11. Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics* **15**(4), 603–619 (2014)
12. Patanè G., Nicolussi F., Krauth A., Colosimo B. M., Dede’ L., Menafoglio A.: *Functional-Ordinal Canonical Correlation Analysis With Application to Data from Bioprinting Sensors*. MOX Report (2024)
13. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. *Ann. Stat.* **23**, 73–102 (1995)
14. Fechner, P., Gauglitz, G., Proll, G.: (2022) Through the looking-glass - recent developments in reflectometry open new possibilities for biosensor applications. *TrAC, Trends Anal. Chem.* **156**, 116708 (2022)

15. Gauglitz, G.: Direct optical detection in bioanalysis: an update. *Anal. Bioanal. Chem.* **398**, 2363–2372 (2010)
16. Hutterer, J., Proll, G., Fechner, P., et al.: Parallelized label-free monitoring of cell adhesion on extracellular matrix proteins measured by single colour reflectometry. *Anal. Bioanal. Chem.* **414**, 575–585 (2022)



Reurbanize Small Towns by International Migration: A Possible Trade off a Possible Trade off a Possible Trade off

Michela C. Pellicani¹(✉), Antonella Bianchino², Monica Carbonara²,
Antonella Rotondo², Roberto A. Palumbo^{1,2}, and Massimo Iaquina¹

¹ University of Bari, Bari, Italy

michelacamilla.pellicani@uniba.it

² Italian National Institute of Statistics, Bari, Italy

Abstract. The picture of Italy in the last ten years, is a demographic black hole in particular if we consider small and micro-towns. This trend will even intensify in the next years compromising the economic development but also the cultural identity, the environmental sustainability, and the supply of fundamental social services like education, health, etc. In this scenario we aim to investigate if this trend could be reversed and with which strategies. At the moment the findings of the scientific literature on these issues are jet quite scarce.

In this study we started from the data produced by the Italian Institute of Statistics (ISTAT) describing, on one hand, the past demographic evolution (in the last 20 years) and, on the other, the projected evolution to 2041. Still from ISTAT, we used the data on internal and international migration in Italy at the municipality level. Basing on some indicators like the migration index (net migration/total migration) we measured the attraction/repulsion ability of the so-called *Inner Areas*. Finally, we combined the demographic approach to the geographical one using territory-based models to detect and bring out the main dynamics of the analysed phenomenon.

Keywords: Depopulation · Inner Areas · Migration

1 Topic

The picture of Italy, in particular, in the last ten years, is a demographic black hole. The small and micro-towns (less than 500 inhabitants), more than others, are losing on average 11.6% of their population; the towns with 500–1,000 inhabitants 9.0% and the towns with 1,000–3,000 inhabitants 7.0% (ISTAT, 2022).

Following the ISTAT projections (ISTAT, 2023a), this trend will even intensify in the next years compromising the economic development but also the cultural identity, the environmental sustainability and the supply of fundamental social services like education, health, etc.. Southern Italy is particularly affected by this phenomenon induced by a consistent drop of births combined with a resurgence of important levels of young adults' emigration (ISTAT, 2023b).

All the different Italian towns/cities are classified in the following six typologies: Urban Hub, Inter-municipal Hub, Belt, Intermediate, Peripheral, and Ultra-peripheral.

The municipalities Hub (Polo) and Inter-municipal Hub have been identified as those that have direct access to the three main typologies of services: education, health, and transport facilities. Depending on the distance from these municipalities, the Belt, Intermediate, Peripheral and Ultra-peripheral municipalities have been identified¹. In 2022, a new mapping of the municipalities was created by the Department for Cohesion Policies and ISTAT², reshaping the definition of essential services to identify the Inter-municipal Hubs and Hubs, and reviewing the minimum distance from the Hubs, functional to the classification of municipalities in other clusters.

A large part of the Italian territory is characterised by a spatial organisation based on minor urban centres of small dimension and limited access to essential services. Nevertheless, these minor urban centres are crucial for the whole national territory in terms of socio-economic development but also in terms of preservation of the cultural identity, of the environmental sustainability, etc.. Hubs, Inter-municipal Hubs and Belt cities and towns are included in the *Centres*. Intermediate, Peripheral, and Ultra-peripheral towns represent the so-called *Aree Interne* (3,834 towns, which represent the 48.5% of the total number of the Italian towns/cities).

These Inner Areas are concentrated in Southern Italy (1,718 towns; 44.8% of the total) and their number has been increasing in the last years. They represent the 67.4% of the Southern Italy municipalities. The most concerned regions are Basilicata, Sicily, Molise, and Sardinia (more than 70%). An even worst situation concerns, in particular, the Ultra-peripheral Southern towns, which are the 59.9% of the total.

The Inner Areas are characterised by a demographic decline that is even more pronounced the longer the distance from the essential services: - 3.4% for the Intermediate towns, - 5.2% for the Peripheral, and - 6.1% for the Ultra-peripheral ones (ISTAT, 2022). In other words, the first one hundred towns included in the Inner Areas marked by the strongest demographic decrease between 2001 et 2020 lost the 40.9% of their population (from 90,188 to 53,314 inhabitants) facing a very serious risk of vanishing.

Of course, part of this regressive evolution is due to the accentuated population ageing affecting the whole Italian territory. At the same time, if for the Hubs and the Inter-municipal Hubs this trend is mitigated by immigration (internal and international), for the Inner Areas both dynamics (natural and migratory) are working in the same direction combining in a vicious spiral.

2 Data and Methods

For our analysis we used the data produced by the Italian Institute of Statistics (ISTAT) describing the past evolution (in the last 20 years) (<https://demo.istat.it/>) and projecting the future evolution to 2041 (<https://www.istat.it/it/archivio/273725>). Still from ISTAT,

¹ Travelling distance on the road to Hubs or Inter-municipal Hubs equal to 27.7 min for Belt; 40.9 for Intermediate; 66.9 min for Peripheral, and > 66.9 min for Ultra-peripheral municipalities.

² Dipartimento per le Politiche di Coesione della Presidenza del Consiglio dei Ministri (2022). Aggiornamento 2020 della mappa delle Aree Interne. Nota tecnica NUVAP.

we used data on internal and international migration in Italy produced thanks to “La rilevazione sulle iscrizioni e cancellazioni anagrafiche per trasferimento di residenza” that provides the basic information for all analyses on internal and international migratory flows. It represents a useful source of data because it allows to measure the intensity and the direction of the flows as well as the main socio-demographic characteristics of those who move their residence. On the internal front, the data used allow to study over time the attractiveness of some areas of the territory and to monitor the consistency of migratory movements and consequently the impact that such transfers have on the population residing in the areas of origin and of destination of the flows. On the international front, the analysis of migratory flows disaggregated by nationality allows to evaluate, in particular, the trend of foreign immigration to Italy and the emigration of Italians abroad (ISTAT, 2023c).

We enriched our analysis using some indicators like the migration index (net migration/total migration) to measure the attraction/repulsion ability of the different typologies of towns.

Furthermore, we combined the demographic approach to the geographical one using territory-based models to detect and bring out the main dynamics of the analysed phenomenon. We produced, as well, a synthesis using a GIS-based analysis which, applied to the social sciences, effectively contribute to highlight the different aspects linked to the territory, the geographical position, and the spatial relationship and to capture, at the same time, many important connections using a multi-disciplinary and multi-factorial approach. The territorial reality of small municipalities and its phenomena are a complex system in which population, cultural factors, energy resources, production factors, physical, and non-physical infrastructures coexist. Plus, they are the territorial basis to know the current situation, its historical transformation, and its evolutionary scenarios. The Geographic Information System (GIS) is particularly useful in this analysis because it is based on the logical model of georeferencing of data and indicators and on the different information layers that overlap and interact with each other while maintaining their own level.

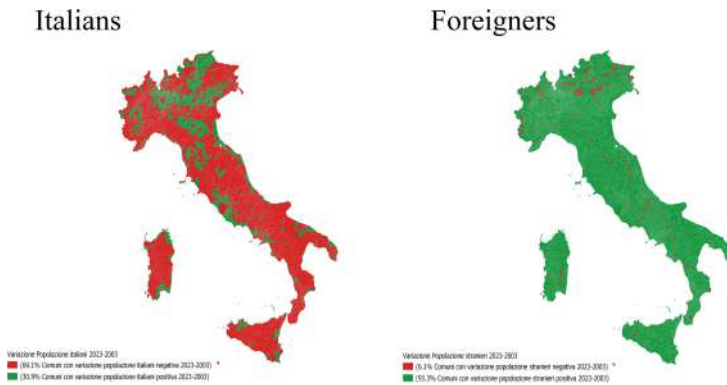
3 Preliminary Findings

Between 2003 and 2023, national demographic growth (+3.2%) is even more pronounced in the Centres (+4.8%), while Inner Areas lose population overall (−2.2%). Specifically, it is the more marginal municipalities within Inner Areas, those Peripheral (−5.4%) and Ultra-peripheral (−9.8%), that decline the most. The Inner Areas municipalities with the greatest population contraction in relative terms are mainly located in Basilicata, Calabria, Sicily, central Sardinia, northern Friuli-Venezia-Giulia.

The depopulation trend primarily stems from the consistently negative natural population movement. The natural balance is negative almost everywhere and in all the years considered. As of January 1, 2023, the demographic decline stands at −5.5%, assuming greater relevance in municipalities in Inner Areas (−6.7%), especially in Peripheral (−7.3%) and Ultra-peripheral (−8.2%) municipalities.

Regarding international mobility, if we consider the entire national territory, in recent years, there has been a substantial stabilisation of the resident foreign population (about 5

million), which remains significantly younger than the Italian population. The population under 10 years old is proportionally larger among foreign citizens (11.7% of its total residents) compared to Italian citizens (7.4%). The foreign population pyramid is further broader even in the 20–40 age group, before starting to narrow in the 40–49 age group and decrease, compared to the Italian population pyramid, starting from the 50–59 age group (Map 1).



Map 1. Variation 2003–2023 in the Italian and foreigner resident populations. Source: own elaboration on ISTAT data

To verify the attractiveness or expulsive force of various territories, the migration index (ratio between the migration balance and the total of registrations and cancellations) is considered. It has been calculated both for the Centres and the Inner Areas.

The index values are generally quite moderate, indicating a substantial parity between the number of registrations and cancellations. Considering the anomaly characterising the 2020 data, it can be inferred that there is a greater tendency to abandon places of residence the farther one is from a service centre (the index values are positive for Centres and negative for Inner Areas) and that this tendency increases over time (ISTAT, 2022).

Finally, from a productive standpoint, the Inner Areas confirm their marginal nature compared to the rest of the country. They absorb only 17.3% of employees in local units of industry and services, while the remaining share is located in the Centres (82.7%). The gap is even more evident when referring to value-added shares: 85.7% comes from local units of businesses located in the Centres, and only 14.3% in the Inner Areas. In particular, the smallest shares are observed precisely in Ultra-peripheral municipalities, with 0.9% of employees and 0.6% of value added.

Even with reference to the taxable income reported in tax returns from all types of taxpayers, the weight of Inner Areas is quite limited: only 17.7% of the total (year 2021, the latest available), compared to 82.3% for Centres. The average taxable income per taxpayer is €17,148 for Inner Areas and reaches €21,723 for Centres. However, it is interesting to note that over the years (2012–2021), this per capita value has experienced a greater increase in Inner Areas (+7.4%) compared to Centres (+4.5%).

4 Some Conclusions

In such a scenario, our questions are: Could this trend be reversed? By which strategies?

An important example of a policy aimed at improving the living conditions of the population residing in areas at risk of marginalisation is the National Strategy of Inner Areas (SNAI) that finds its regulatory reference in the 2014 National Reform Program (PNR) and is defined in the 2014–2020 Partnership Agreement³.

The SNAI has been launched in 2012 and it has been financed with more than 550 million euros dedicated to some local development projects in 72 areas of intervention regrouping 1 thousand towns and 2 million inhabitants.

The policy approach was place based and it aimed to directly involve the local institutional actors and associations to maximise the effectiveness of the intervention.

However, the insufficient results pushed us to another question. Considering the population ageing and the depopulation of small-medium towns on one hand, and the numerous problems related to the integration process of the migrants on the other hand, could it be possible to trigger a positive trade off that could be of mutual interest both for the Inner Areas (repopulation) and for the foreign population in Italy (a successful integration at least for a part of it)?

The entrepreneurial dimension of the small towns, as a matter of fact, is more oriented to the local know how and the prevalence of the crafts and agricultural activities is positively correlated to the territorial peripherality. At the same time, a large number of immigrants are characterised by an important know how for crafts and agriculture activities associated with high entrepreneurial skills.

Basing on these attitudes and competences, a possible path to deeper explore is, therefore, to promote immigrants' active integration policies to repopulate and revitalise small and medium towns through the development of new enterprises in the crafts, agriculture, and tourism sectors.

References

- Calvaresi, C.: Una strategia nazionale per le aree interne: diritti di cittadinanza e sviluppo locale. *TERRITORIO* **74**, 78–79 (2015). <https://doi.org/10.3280/TR2015-074013>
- Dipartimento per le Politiche di Coesione e per il Sud della Presidenza del Consiglio dei Ministri. <https://www.politichecoesione.governo.it/it/politica-di-coesione/strategie-tematiche-e-territoriali/strategie-territoriali/strategia-nazionale-aree-interne-snai/le-aree-interne-2021-2027/mappa-aree-interne-2020/>
- Eimermann, M., Carson, D.A.: European lifestyle migrant entrepreneurs and their business network in swedish sparsely populated areas. In: Kordel, S., Weidinger, T., Jelen, I. (eds.), *Processes of immigration in rural Europe: the status quo, implications and development strategies*, pp. 243–269. Cambridge Scholars Publishing (2018)
- Formez, <http://focus.formez.it/content/strategia-nazionale-aree-interne-e-nuovi-assetti-istituzionali>

³ Dipartimento per le Politiche di Sviluppo della Presidenza del Consiglio dei Ministri (2013). *Strategia nazionale per le Aree Interne: definizione, obiettivi, strumenti e governance – Documento tecnico collegato alla bozza di Accordo di Partenariato trasmessa alla CE il 9 dicembre 2013.*

- Fromentin, J.: *Des configurations géographiques en mouvement: l'immigration dans les espaces peu denses français*. European J. Geography Espace, Société, Territoire (2022). <https://doi.org/10.4000/cybergeo.39893>
- Hubbard, P.: 'Inappropriate and incongruous': opposition to asylum centres in the English countryside. J. Rural. Stud. **21**(1), 3–17 (2005). <https://doi.org/10.1016/j.jrurstud.2004.08.004>
- ISTAT: La geografia delle aree interne nel 2020: vasti territori tra potenzialità e debolezze (2022)
- ISTAT: Previsioni della popolazione residente e delle famiglie – Base 1/1/2022, Statistiche Report (2023a)
- ISTAT: Rapporto annuale 2023. La situazione del Paese (2023b)
- ISTAT: Migrazioni interne e internazionali della popolazione residente. Anno 2021, Statistiche Report (2023)
- Jentsch, B., de Lima, P., Brian, M.: Migrant workers in rural Scotland: “going to the middle of nowhere.” Int. J. Multicult. Soc. **9**(1), 35–53 (2007)
- Lymperopoulou, K.: Immigration and ethnic diversity in England and Wales examined through an area classification framework. J. Int. Migr. Integr. **21**(3), 829–846 (2019). <https://doi.org/10.1007/s12134-019-00678-9>
- Massey, D.S. (ed.): *New Faces in New Places: The Changing Geography of American Immigration*, p. 370. Russell Sage Foundation, New York (N.Y.) (2008)
- McAreavey, R.: *New Immigration Destinations: Migrating to Rural and Peripheral Areas*, p. 302. Routledge (2017). <https://doi.org/10.1201/9780415540056>
- Morén-Alegret, R., Wladyka, D.: *International Immigration, Integration and Sustainability in Small towns and Villages: Socio-Territorial Challenges in Rural and Semi-rural Europe*, p. 380. Palgrave Macmillan, London (2020)
- Neal, S.: *Rural Identities: Ethnicity and Community in the Contemporary English Countryside*, p. 159. Routledge (2016). <https://doi.org/10.4324/9781315607177>
- Nelson, L., Hiemstra, N.: Latino immigrants and the renegotiation of place and belonging in small town America. Soc. Cult. Geogr. **9**(3), 319–342 (2008). <https://doi.org/10.1080/14649360801990538>
- Woods, M.: Engaging the global countryside: globalization, hybridity and the reconstitution of rural place. Prog. Hum. Geogr. **31**(4), 485–507 (2007). <https://doi.org/10.1177/0309132507079503>



Network Analysis: An Application to EU Geographical Indications

Alfonso Peluso^{1(✉)}, Antonella Bianchino^{2,3}, and Agata Maria Madia Carucci²

¹ Department of Economics and Statistics, University of Salerno, Fisciano, (SA), Italy

alpeluso@unisa.it

² Ufficio Territoriale Basilicata, 85100 Basilicata, Potenza, Italy

³ Ufficio Territoriale Area Sud, Istituto Nazionale di Statistica, Basilicata, Italy

Abstract. This work aims at displaying and analyzing the relationship between PDO and PGI products and the territories they are produced in. It does so through the use of network science, depicting the relationships as a bipartite network. This approach proves useful in displaying in an intuitive manner different features of the products and the territories, given its flexibility and broad range of applicability to different measures. To outline its usefulness, an application is performed on the products and territories of the Italian Puglia region to show some of the possible applications.

Keywords: Network Science · PDO products · bipartite networks · monpartite networks

1 Introduction

Network science [6], which builds upon concepts drawn from graph theory and statistical physics, has been applied to many research topics in various fields such as Economics, Biology, Computer science and Social sciences. It proves useful in many areas as a tool for representing and analyzing the relationships among observational units with applications ranging from visualization to numerical representations of the structure of the network. For what concerns territories, it has recently been applied to tourism in various ways [1,2] and to the analysis of the agricultural sector [3–5]. This paper gives a novel perspective about the application of network science to the relationships among Protected Designation of Origin (PDO) and Protected geographical indication (PGI) products and territories, with an analysis focused on the Italian Puglia region showing one of the possible applications.

2 Methodology

Before delving deeper into the theoretical interpretation, the foundational concepts will be defined. A graph is a set of vertices (nodes) and links (edges),

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 416–421, 2025.

https://doi.org/10.1007/978-3-031-64447-4_70

defined as $G(V, E)$ with V indicating the vertex (node) set and E the links (edges) set respectively. Graphs can have different definitions based on the type of edges they have (directed-undirected, weighted-unweighted). Graphs can also have different types of nodes and, usually, when there is more than one type of nodes the edges connect only nodes of different types. These kind of graphs are defined as multipartite graphs. In the particular case of two types of nodes they are defined as bipartite graphs. Given that most of the theoretical results are defined in the context of monopartite graphs, the standard way of dealing with bipartite graphs is to work on monopartite projections of these graphs, i.e. if node A and node B are connected to node α in the bipartite graph, then a monopartite projection can be obtained by connecting node A and B in a monopartite graph. Both bipartite graphs and monopartite projections will be used in the analysis.

2.1 Product-Territory Bipartite Network

The bipartite network structure proves useful in displaying the product-territory relationships as a bipartite graph to get an intuitive representation of the presence of PDO (PGI) products on territories. It can be seen from Fig. 1 that the network gives insights about the spread of olive oil production across the region.

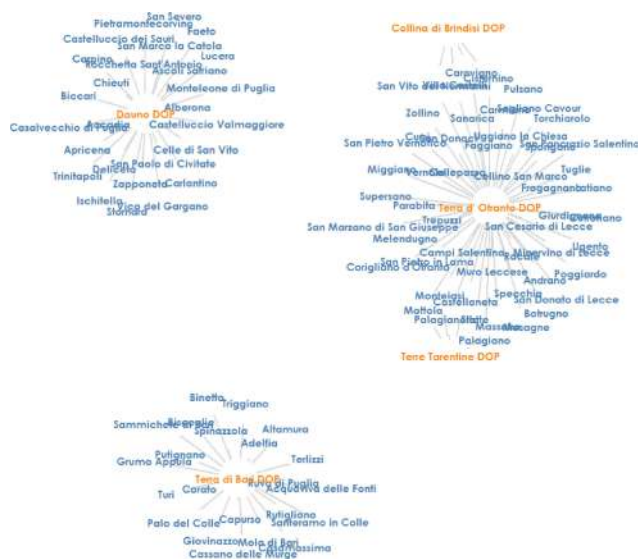


Fig. 1. Product-Territory bipartite network for the PDO oils of the Italian Puglia region. To enhance readability, the plot was generated by randomly selecting 100 municipal units from the entire dataframe and displaying the associated bipartite subnetwork. Product nodes are depicted in orange, while nodes representing territories are displayed in blue.

2.2 Product-Product Monopartite Projection

The product-product monopartite projection, obtained connecting the products connected to the same territories in the bipartite network above, with edge weights corresponding to the number of common territories between products can be used to visualize which products are connected to the same territories (Fig. 3) and to gain insights from some calculated network statistics such as the node degree, a measure of connectedness of each of the nodes in the graph, which can be interpreted as a measure of the co-presence of a product together with another on certain territories or the average degree which, in this case, can be regarded as an indication of the average number of territories shared by two or more products.

2.3 Territory-Territory Monopartite Projection

The territory-territory monopartite network (Fig. 2) is obtained by linking the territories connected to the same products in the bipartite network described above. It is useful in visualizing the relationships between territories and to get some useful information from network statistics. In particular the degree of the nodes in this network can be used to identify hubs (territories with a high number of connections) and peripheral nodes (territories with a low number of connections), the average degree gives an indication of the average number of territories that share one or more products and so can be used to analyze the spread of PDO and PGI products across the territories of a region or country.



Fig. 2. Territory-Territory monopartite projection of the bipartite network of Fig. 1.

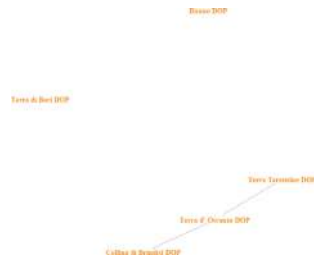


Fig. 3. Product-Product monopartite projection of the bipartite network of Fig. 1.

3 Networks from Data

The analysis is performed using data sourced from the MIPAAF website, gathered in an unstructured manner as shape files (.geojson) delineating production

zones for each PDO (PGI) product and then transformed in a structured way by the author(s). This process led to the creation of a tabular dataset with territories as observations and products as columns and 0–1 entries to indicate the presence of a certain product in a certain territory and provided a way of representing these relationships as a bipartite network connecting products to production zones. The products are split into categories, according to the EU classifications, the categories are: Wine, Fruit, vegetables and cereals Cheeses, Oils and fats, Bakery products, pastries, other products of animal origin.

The analyses will be performed on the oil network only (just PDO oils) for ease of exposition, a more detailed analysis would consider the whole products network, the whole wines network and the complete network (including all products and all wines). Additional data used in the analysis are product-specific and territory-specific data from ISTAT [8] indicating the number of operators and processors for each product.

4 Analysis

The analysis is performed on data from the Italian Puglia region to give some insights on the structure of the various networks described above and displaying additional information about the number of operators for products belonging to the Oils and fats category and the corresponding territories.

Table 1. Number of Operators and processors for each Oils and Fats PDO product

PDO Oil Product	N. of Opertors and Processors
Collina di Brindisi DOP	162
Dauno DOP	120
Terra d' Otranto DOP	372
Terra di Bari DOP	2756
Terre Tarentine DOP	98

Table 1 displays the number of producers and processors for the PDO Oils and fats of the Puglia region while Table 2 shows the top five territories by number of operators in the PDO and PGI products of the Puglia Region. These two dataframes measure the same thing on two different observational units, but can be displayed together through a bipartite network. As Fig. 4 shows the PDO Oils and Fats bipartite network with the node sizes proportional to the number of operators for both products and territories. It can be seen that the products differ for both the structure and the diffusion across the territory, in particular the Terre di Bari is connected to fewer but bigger nodes while the Terra d' Otranto, being connected to more territories has a wider extension but a smaller average number of operators, lastly the Dauno has a smaller extension with respect to both the Terre di Bari and the Terra d' Otranto and is more

Table 2. Number of operators and processors in Oils and Fats PDO products for each of the top 5 territories

Territories	N. of Opertors and Processors
Andria	645
Bitonto	457
Ruvo di Puglia	148
Barletta	132
Altamura	123

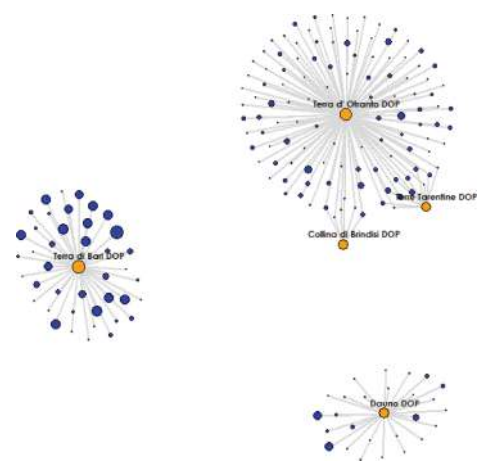


Fig. 4. Product-Territory bipartite network for the PDO oils and fats of the Italian Puglia region. The vertex size is proportional to the number of producers and processors. The size of the nodes is proportional to logarithm of the number of processors.

heterogeneous for what concerns the average number of operators. The analysis can be extended to the whole network of products to directly analyze the impact of each product on each territorial unit.

5 Conclusions and Future Developments

This work introduced a new framework for the analysis of the impact that PDO and PGI products have on the territories. Representing the relationships between products as a bipartite network and extracting and analyzing the two monopartite projections of the bipartite network can give many insights on the spread of quality food products across a certain territory. Furthermore, by quantitatively analyzing the two monopartite networks, a new set of indicators based on network statistics can be derived based on the accessory data used for the analysis. The main feature is the ability to easily display the connection between

products and territorial units, the main limitations are observed with respect to the number of units that can be displayed in the network. Possible applications could be relative to the direct analysis of the economic impact of a product on the territories that can produce it, to analyze the impact of PDO products on tourism and farm holidays development or to other kinds of data related to the products under analysis.

References

1. Baggio, R., Scott, N., Cooper, C.: Network science: a review focused on tourism. *Ann. Tour. Res.* **37**(3), 802–827 (2010)
2. Van der Zee, E., Vanneste, D.: Tourism networks unravelled; a review of the literature on networks in tourism management studies. *Tourism Manag. Perspectives* **15**, 46–56 (2015)
3. Hernández, M.L.H., et al.: An application of network theory to territorial analysis: the case of Yaqui Valley networks (2006 and 2016). *Revista de la Facultad de Ciencias Agrarias UNCuyo* **50**(2), 311–329 (2018)
4. Noguera-Talavera, Á, Salmerón, F., Reyes-Sánchez, N.: Bases teórico-metodológicas para el diseño de sistemas agroecológicos. *Revista de la Facultad de Ciencias Agrarias. Universidad Nacional de Cuyo* **51**(1), 273–293 (2019)
5. Bezard, M., et al.: Co-designing innovative plantain cropping systems to support the diversity of agroecological pathways in Guadeloupe. *Agronomy Sustainable Develop.* **43**(2), 28 (2023)
6. Newman, M.: *Networks*. Oxford University Press (2018)
7. Prodotti DOP e IGP, MIPAAF. <https://dopigp.politicheagricole.it/>
8. Prodotti agroalimentari di qualità DOP/IGP e STG, ISTAT Data, <https://esploradati.istat.it>



Immigrants' Children and Socio-economic Achievement: An Analysis Across Western Europe

Giorgio Piccitto^(✉)

Department of Sociology and Social Research, University of Milano-Bicocca, Milano, Italy
giorgio.piccitto@unimib.it

Abstract. The socio-economic inclusion of immigrants' children is a marker to evaluate the ethnic-based mechanisms of transmission of social inequality. Anyway, little is known about these mechanisms and how they operate in countries that only recently arose as migrant destinations, like the Southern European ones. This represents an important gap in the literature, since Southern European countries present peculiar characteristics in terms of patterns of socio-economic assimilation of first-generation migrants, and these peculiarities may also affect their children's assimilation trajectories. This work aims to fill this gap and provides, for the first time, a comprehensive overview of the migrants' socio-economic achievement in 9 Western European countries, accounting for the different migrant generations and for their stratification-relevant characteristics. To this purpose, multivariate statistical techniques are applied to the individual data of two Eurostat "ad hoc modules" (2014, 2021) EU Labour Force Survey. Results suggest cross-countries heterogeneities in the patterns under scrutiny.

Keywords: migrants · second generation · ethnic penalty · social origin · socio-economic achievement

1 Motivation

Children of immigrants are becoming an increasingly relevant component of European societies, and their study is gaining momentum and broadening its range and scope. Of particular concern is their socio-economic inclusion in the host labor market, as a way to evaluate from a long-term perspective the process of ethnic stratification in the receiving country and the mechanisms of transmission of social inequality (Zhou and Bankston, 2016). Relatedly, the literature has largely recognized the relevance of the individual socio-economic background of origin as a driver of the immigrants' children "fate" (Heath et al., 2008; Portes and Zhou, 1993), and that this may operate differently among native individuals and those with a migratory background (Avola and Piccitto, 2020; Panichella et al., 2021; Zuccotti and Platt, 2023). Still, this issue has, due to absence of data, understudied in a cross-country perspective. Still, this issue has been understudied in a cross country perspective, due to absence of data.

Also, the considerable increase of research on socio-economic attainment of immigrants' children has focused on old migration countries, with this topic remaining largely unexplored in the new receiving contexts (Gabielli and Impicciatore, 2022).

This work aims at filling these gaps, by exploiting two Eurostat “ad hoc modules” (2014 and 2021) of the European Union Labour Force Survey (EU-LFS).

Applying multivariate statistical techniques to these data, the present work provides, for the first time, a comprehensive overview of the migrants' labor market achievement in 9 Western European countries, accounting for the different migrant generations and for individual stratification-relevant factors. In this way, it is possible to provide an assessment of the ethnic stratification of migrants belonging to different generations in the European labor markets, accounting for their social class of origin in the analyses.

Starting from the OED triangle (Blau and Duncan, 1967), different links of the interplay between migration status, education, and social origin in shaping the individual labor market outcomes will be inquired (see the theoretical framework in Fig. 1).

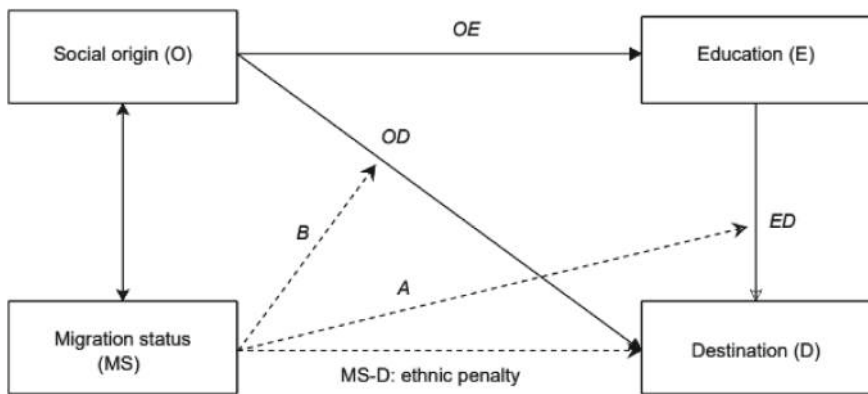


Fig. 1. Theoretical model of the interplay between migration status, education, and social origin in shaping labor market socio-economic outcomes. From Piccitto et al. (2024).

2 Data, Variables and Methods

The two “ad hoc modules” of the European Union Labour Force Survey (EU-LFS) used for the analyses contain key information; not only they include the information on own country of birth, but also on the country of birth of the respondent's father and mother; also, they contain the information on the respondent's year of arrival in the host country. With these variables it is possible to precisely define the migrants' generation. In the analyses, 9 Western European countries are considered: Austria, Belgium, Denmark, France, Germany, Italy, Spain, Sweden, Switzerland.

The sample consists of men and women aged 16-64. Individuals in the sample are distinguished on the basis of their migration status, as follows: a) natives those born in the ‘host’ country with both native parents; b) first-generation migrants from high emigration countries (Hec): those born in a less socio-economically developed countries and arrived

in the ‘host’ country when older than 12 years; c) immigrants’ children: those born in the ‘host’ country with at least one foreign parent and those born abroad (with at least one foreign parent) but migrated in Italy when younger than 7 years.

The analyses estimate the extent to which the social origin (measured through the parental education) moderates the association between the migration status and the probability of having a tertiary education. The results tap into the literature of the immigrants’ children socio-economic inclusion into the labor market from a segmented assimilation perspective (Portes and Zhou, 1993).

Analyses are conducted by means of OLS models. The coefficient of interest is operationalized through an interaction term between migration status and parental education. For sake of parsimony, results are shown only for migrants’ children (see below).

3 Selected Results

Preliminary results underline that there are some differences in the β coefficients quantifying the gap between low-social origin (left part of the graph) and high-social origin (right part of the graph) natives and migrants’ children in the probability of having a tertiary degree. across countries. The analysis underline the existence of different patterns in the considered countries. Such findings may be interpreted in terms of different socio-economic patterns of integration of first-generation migrants, migration histories or different welfare systems of the settlement countries.

Further steps may shed light on how the interplay between migration status, social origin and education in shaping the socio-economic achievement varies distinguishing

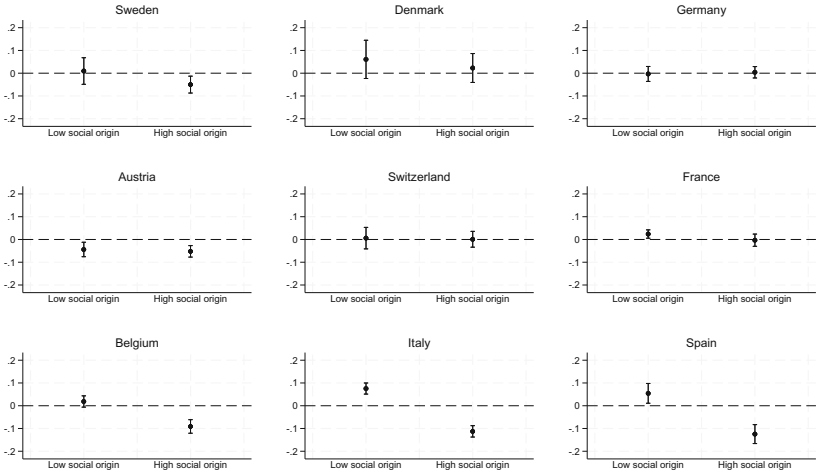


Fig. 2. Difference in the probability of having a tertiary degree among immigrants’ children and natives by individual social origin. Notes: the dots represent the β coefficients based on the ordinary least squares models after controlling for gender, age, year of survey, education. The lines represent 90% confidence intervals.

the parental country of origin; indeed, this has emerged as an important stratifying factor in the fate of immigrants' children (Piccitto, 2023).

References

1. Avola, M., Piccitto, G.: Ethnic penalty and occupational mobility in the Italian labour market. *Ethnicities* **20**(6), 1093–1116 (2020)
2. Blau, P.M., Duncan, O.D.: *The American Occupational Structure*. John Wiley & Sons Inc (1967)
3. Gabrielli, G., Impicciatore, R.: Breaking down the barriers: educational paths, labour market outcomes and wellbeing of children of immigrants. *J. Ethnic Migr. Stud.* **48**(10), 2305–2323 (2022)
4. Heath, A.F., Rethon, C., Kilpi, E.: The second generation in Western Europe: education, unemployment, and occupational attainment. *Ann. Rev. Soc.* **34**, 211–235 (2008)
5. Panichella, N., Avola, M., Piccitto, G.: Migration, class attainment and social mobility: An analysis of migrants' socio-economic integration in Italy. *Eur. Soc. Rev.* **37**(6), 883–898 (2021)
6. Piccitto, G.: The later, the better? The ethnic penalty on labor market achievement by migrant generation: evidence from Italy. *Sociologia del Lavoro* **166**(2), 79–101 (2023)
7. Piccitto, G., Avola, M., Panichella, N.: Migration, social stratification, and labor market attainment: an analysis of the ethnic penalty in 12 Western European countries. *Int. J. Comp. Soc.* 00207152241246166 (2024)
8. Portes, A., Zhou, M.: The new second generation: segmented assimilation and its variants. *Ann. Am. Acad. Polit. Soc. Sci.* **530**(1), 74–96 (1993)
9. Zhou, M., Bankston, C.L., III.: *The Rise of the New Second Generation*. John Wiley & Sons (2016)
10. Zuccotti, C.V., Platt, L.: The paradoxical role of social class background in the educational and labour market outcomes of the children of immigrants in the UK. *The British J. Soc.* **74**(4), 733–754 (2023)



Unlocking Urban Insights: Leveraging Open Data for Governance and Research

Giorgio Piccitto^(✉), Domingo Scisci, and Sonia Stefanizzi

Department of Sociology and Social Research, University of Milano-Bicocca, Milano, Italy
giorgio.piccitto@unimib.it

Abstract. Open data is a tool with enormous potential, both for implementing open government and participation practices and for improving the effectiveness of academic research, by increasing data transparency, accessibility, cost-saving and reproducibility (Boulton et al., 2011). Unfortunately, several impediments have been recognised preventing the open data process from fully realizing its positive effect (Zuiderwijk et al. 2021).

The “Open Government Data” project aims to extract value from administrative data and provide the scientific community with new tools to understand the governance processes of local institutions. Starting from the open data provided by some Italian cities and regions, the project aims to create new ready-to-use research data to be made available to all researchers, adopting a more transparent and collaborative methodological standard and process (the “code-first” approach) (Gentzkow and Shapiro, 2014).

Keywords: open data · cities · governance · data transparency

1 Open Data for Governance and Research

Open data is a tool with enormous potential (Gurstein 2011; Krotoski 2012), both to implement open government and participation practices, and to improve the effectiveness of academic research, by increasing data transparency, accessibility, cost-saving and reproducibility (Boulton et al., 2011).

Such data becomes particularly crucial in understanding new trends in cities’ governance, grounded in the current general trend towards data-driven decision-making (Neves et al., 2020). However, in this scenario of hectic development of digital technologies and information-intensive productive processes, several impediments have been recognised that prevent the open data process from fully realising its positive effects (Zuiderwijk et al. 2021). In particular, it has been recognised the existence of technological barriers to open data adoption, including data quality, accessibility, compatibility, credibility, processability, and lack of standards (Hossain et al. 2016).

2 The “Open Government Data” Project

In this context, we will present the “Open Government Data” project, which aims to extract value from administrative data and provide the scientific community with new tools to understand the governance processes of local institutions. Starting from the open data provided by some Italian cities and regions, the project aims to create new ready-to-use research data to be made available to all researchers.

The working method is based on two specific aspects.

Firstly, typical data curation practices will be applied, such as using metadata, documentation and distribution standards that will guarantee the FAIRness (Briney, 2015; Wilkinson et al., 2016) of the data.

Secondly, the “code-first” approach will be adopted, which consists in focusing attention and work on the code that generates the data, as opposed to the data themselves, facilitating the activities of documenting the operations performed, in a transparent and collaborative manner (Gentzkow and Shapiro, 2014). This approach allows great improvements in terms of reproducibility and accountability, since it keeps track of what has changed, when, by whom and why.

Consistently with this strategy, the entire data life-cycle from raw to public data has been managed following the typical software development workflow. First, a repository managed by a version control software (Git) was created, which ensured the tracking of all changes and made collaboration easier. The code was hosted in a repository manager (GitLab) that allowed the research team to track issues, bugs and requests collaboratively, linking issues to the code used to resolve them.

This procedure enabled the rapid and frequent release of data to the research team, still in the field phase, following an agile-like methodology. Such DevOps-oriented approach also allowed the definition of standardized procedures for setting up the survey instruments (questionnaires), technical documentation and metadata needed for long-term preservation.

This second aspect – which consists of treating the entire data lifecycle according to the typical software development workflow – represents one of the most original aspects of the project, which will guarantee its sustainability even after its completion. The strengths and weaknesses of this approach will be discussed, particularly highlighting the need for more IT skills in the social sciences and for new professional figures who can integrate technical skills and domain-specific knowledge.

3 Objectives, Areas of Intervention and Expected Outcomes

The project aims to define and propose a new rationale to produce and interpret the Public Administration Open Data (PAOD), to overcome all those accessibility issues that prevent the full realization of the Open Science (OS) paradigm and that reproduce structures of social inequalities (on the basis of individual material and competencies resources).

Specifically, three broad goals are defined: 1) to establish the Public Administration Open Data Observatory; 2) to develop new research lines, inspired by and coherent with the new proposed ‘open’ and ‘accessible’ way of conceiving data; 3) to create the

Innovative Teaching Lab, through which will be developed some good practices in terms of multimedia production and MOOC (Massive, Open, Online Courses) provision.

The project is explicitly targeted to increase social and economic efficiency and sustainability in local social contexts (Vega-Tinoco et al., 2022). Indeed, the availability of open data contributes to the citizens' empowerment and mobilization and can be used as a benchmark to ask for public policies and interventions and to improve the effectiveness of evidence-based intervention.

Several promising concrete applications are expected, allowing to overcome some of the most known problem related to the open data quality and accessibility, especially in terms of completeness and accurateness of the information. One objective will be the systematization of the information at a sub-city level. Concretely, it will be created a database of individual socio-demographic characteristics at the 'town halls' (*Municipi*) level in Milan, that would be freely consultable.

This database will trigger new avenues also in Academia, permitting research at a very fine-grained territorial level. Such datasets, indeed, could be used to pass from the aggregated to the micro level, recreating a dataset at the individual level by weighting each combination of the combined variables. This method, recently applied (with Eurostat data) in a cross-country study (Ortensi et al., 2024), could be adopted in the study of urban phenomena allowing the use of multivariate analyses techniques.

References

- Boulton, G., Rawlins, M., Vallance, P., Walport, M.: Science as a public enterprise: the case for open data. *The Lancet* **377**(9778), 1633–1635 (2011)
- Briney, K.: Data Management for Researchers: Organize, Maintain and Share your Data for Research Success. Pelagic Publishing Ltd. (2015)
- Gentzkow, M., Shapiro, J.M.: Code and Data for the Social Sciences: A Practitioner's Guide. University of Chicago, Mimeo (2014)
- Gurstein, M.B.: Open data: Empowering the empowered or effective data use for everyone? *First Monday* **16**(2) (2011)
- Hossain, M.A., Dwivedi, Y.K., Rana, N.P.: State-of-the-art in open data research: Insights from existing literature and a research agenda. *J. Organ. Comput. Electron. Commer.* **26**(1–2), 14–40 (2016)
- Krotoski, A.K.: Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise. *Insights: the UKSG J.* **25**(1), 28–32 (2012)
- Neves, F.T., de Castro Neto, M., Aparicio, M.: The impacts of open data initiatives on smart cities: a framework for evaluation and monitoring. *Cities* **106**, 102860 (2020)
- Ortensi, L., Piccitto, G., Baiocchi, S.: A 'nuanced female advantage'. A gendered analysis of decisions on asylum applications in Italy from 2008 to 2022. *Genus* **80**, 1–22 (2024)
- Vega-Tinoco, A., Gil-Lacruz, A.I., Gil-Lacruz, M.: Civic participation as a promoter of well-being: comparative analysis among European countries. *Soc. Indic. Res.* **164**(1), 217–237 (2022)
- Wilkinson, M., Dumontier, M., Aalbersberg, I., et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016)
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R.S.: Socio-technical impediments of open data. *Electron. J. e-Gov.* **10**(2), 156–172 (2012)



Tourism Sustainability: A Statistical Study in Friuli Venezia Giulia Municipalities

Gianfranco Piscopo¹, Gabriella Schoier², Massimiliano Giacalone³(✉),
and Maria Longobardi¹

¹ University of Naples “Federico II”, 80138 Napoli, Italy
{gianfranco.piscopo, maria.longobardi}@unina.it

² University of Trieste, 34127 Trieste, Italy
gabriella.schoier@deams.units.it

³ University of Campania “L. Vanvitelli”, 81043 Capua, Italy
massimiliano.giacalone@unicampania.it

Abstract. The aim of this paper is to consider some indicators for the analysis of tourism in the region Friuli - Venezia Giulia and to summarize the information contained in these indices using a multivariate method: Principal Component Analysis. With this analysis we have taken into account the demand, the supply and the impact of the tourism sector on the environment of the municipalities of the region Friuli - Venezia Giulia. The analysis provides the level of “stress” of the territory, referring to predetermined territorial districts and to annual official data, and contributes to the assessment of tourist sustainability in the territory.

Keywords: Tourism Sustainability · Indicators · Principal Components Analysis

1 Introduction

The concept of sustainability, born in the environmental field, is widely spread to define a “balance between meeting current needs without compromising the possibility of future generations to make up for their” (Brundtland report, 1987). This concept has subsequently been extended to other areas of human activities including tourism, where it refers to many aspects like the respect and protection of the environment, culture, traditions and socio-cultural resources of the local population. The process of integrating the previously four aspect of sustainability is complex and depends on several factors, such as the size of the company, its business nature, and its corporate culture. In this perspective, the company’s objective of sustainability should be translated into a four responsibility, in which the evaluation of the company’s results would depend on the synthesis of economic performance with its environmental, cultural and social impact [9]. In particular, as regards the problem of explaining tourism sustainability we have considered some indicators. These indicators can help in describing the phenomenon of tourism, which is strictly connected with the territories because they can spot situations concerning different places [6].

In this paper, we address the problem of monitoring sustainable tourism in the municipalities of Friuli-Venezia Giulia (F-VG) [8].

The data have been selected from the ISTAT database [7], the regional WebTur database [14] and the regional ARPA database [1]. In the uncertain climate of 2020, with its demographic repercussions on mortality, birth rates and marriages, and with the unknowns on employment, we have however been working on compensations, particularly for the accommodation and catering sector, and above all we have begun to look beyond, in search of a strategy that will make strategic sectors - such as tourism, mobility, health and social services, and public administration more sustainable. In order to try to give an answer to this problem this research proposes some selected indicators and try to understand how they can be summarized using principal component analysis [13].

In particular, as regards the problem of explaining tourism sustainability we have considered some indicators. These indicators can help in describing the phenomenon of tourism, which is strictly connected with the territories because they can spot situations concerning different places as that are still under-exploited (despite of potential tourist interest), and areas at the limits of saturation.

2 The Application

In this section, we describe the data and the indicators used. The second part regards the results of the application of principal component analysis to the indicators.

In order to build the indicators data have been selected from the ISTAT database, the regional WebTur database and the regional ARPA database.

The data, contained within the regional database WebTur, come from ISTAT census surveys on “Capacity of accommodation establishments” and on “Movement of customers in accommodation establishments” where the survey units are the accommodation establishments divided between “hotel facilities” and “complementary facilities”.

The first survey records annually the consistency of accommodation facilities in terms of number and capacity of accommodation establishments annually present in the region. The second survey records, for each month and for each municipality, the arrivals and presences of Italian and foreign customers according to the type of accommodation, the category (star rating for hotel accommodation) and according to the origin of the tourist (country of residence in the case of foreigners, region or autonomous province of residence for Italian customers).

The municipalities have been chosen according to the tourist rate published by the office of Planning, strategic planning, monitoring/control management and statistics of the region F-VG [10–12]. The subdivision of the municipalities according to the tourist rate is reported in Fig. 1.

According to the rate of tourism presented in the report [11] 48 municipalities have been chosen: Ampezzo, Aquileia, Arta Terme, Aviano, Barcis, Cavazzo Carnico, Chiusaforte, Claut, Clauzetto, Duino Aurisina, Fogliano Redipuglia, Forgaria Nel Friuli, Forni Avoltri, Forni Di Sopra, Forni Di Sotto, Gorizia, Gradisca d’Isonzo, Grado, Lignano Sabbiadoro, Malborghetto Valbruna, Monfalcone, Monrupino, Montereale Valcellina, Muggia, Ovaro, Palazzolo Della Stella, Palmanova, Paluzza, Polcenigo, Pordenone,



Fig. 1. Municipalities of *F-VG* source own elaboration

Pradamano, Prato Carnico, Preceniccio, Ravascletto, Sagrado, San Dorligo Della Valle, San Giorgio Di Nogaro, Sappada, Sauris, Socchieve, Sutrio, Tarvisio, Trieste, Trivignano Udinese, Udine, Venzone, Vivaro.

3 Indicators and Results

In order to evaluate the weight of infrastructures both on the physical territory and on the population different indices have been proposed: 1) Tourism Spatial Density Index [3]; 2) Defert Tourist Index [2]; 3) Tourist Pressure Index (at time t) [3]; 4) Settlement Florence Index [4]; 5) Weighted Waste Production (for occupied tourist bed) [3]; 6) Tourism Relation Index [3].

At the beginning, we present some descriptive results then we apply principal component analysis. As one can see from Table 1 the means and the standard deviation are very different even in the normalized indices.

At this point we have applied the principal component analysis. The eigenvalues are reported in Table 2. We note that the first CP explain approximately the 58% of the variance of the original variables. Considering two CP the percentage is about the 79%.

In Table 3 the eigenvectors are presented. The correlation among CP1 and CP 2 and the original variables are reported in Table 4 and Fig. 2. The first component (CP 1) can be defined as and on impact of tourism on physical territory, residents and production of waste; while the second component (CP 2) can be viewed as the impact of tourism on residents and pressure on tourist facilities.

Table 1. Some descriptive statistics.

Indices	Mean	Standard deviation
Index_tourism_spatial_density	55.7185855	259.4346581
Index_Defert_tourist	0.1037630	0.1714695
Index_tourist_pressure	80.2898189	91.2754367
Index_settlement_Florence	0.1037630	0.1714695
Index_prod_waste_tourism	691.5248528	2231.1549715
Index_tourist_relations	185.1433270	412.2815891

Table 2. Eigenvalues, percentage of explained variance, cumulative percentage of variance.

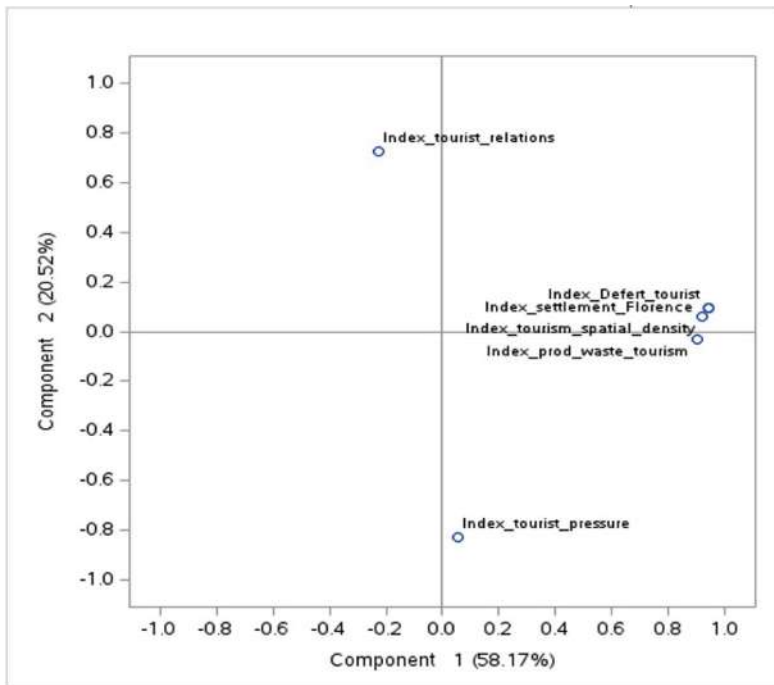
Eigenvalue	Explained variance	Cumulative variance
3.49035370	0.5817	0.5817
1.23127810	0.2052	0.7869
0.82511866	0.1375	0.9245
0.37659567	0.0628	0.9872
0.07665387	0.0128	1.0000
0.00000000	0.0000	1.0000

Table 3. Eigenvectors.

Indices	Principal component_1	Principal component_2
Index_tourism_spatial_density	0.493154	0.053887
Index_Defert_tourist	0.504036	0.086855
Index_tourist_pressure	0.030122	−0.746029
Index_settlement_Florence	0.504036	0.086855
Index_prod_waste_tourism	0.482833	−0.026402
Index_tourist_relations	−0.121080	0.651730

Table 4. Correlations among indices and the first two principal component.

Indices	Principal component_1	Principal component_2
Index_tourism_spatial_density	0.92134	0.05980
Index_Defert_tourist	0.94166	0.09638
Index_tourist_pressure	0.05627	-0.82782
Index_settlement_Florence	0.94166	0.09638
Index_prod_waste_tourism	0.90205	-0.02930
Index_tourist_relations	-0.22621	0.72318

**Fig. 2.** Pattern of component 1 and component 2.

4 Conclusions

The process of integrating the previously four aspect of sustainability is complex and depends on several factors, such as the size of the company, its business nature, and its corporate culture. The purpose of this paper has been to monitor sustainable tourism on the basis of the data relating to municipalities of Friuli - Venezia Giulia.

The data have been selected from the ISTAT database, the regional Web Tur database and the regional ARPA database.

We apply selected indicators to municipalities of F-VG chosen on the basis of the tourism indicator presented by Region F-VG. We have tried to understand how these indicators can be summarized using principal component analysis. The first principal component is determined by imposing the constraint that the variance is maximum under the condition that the eigenvector is normalized. The other principal components are still obtained as linear combinations of the initial variables, they are uncorrelated and explain the maximum variability once the variability of the previous principal components has been removed [5].

The results show that CP 1 can be defined as and on impact of tourism on physical territory, residents and production of waste; while CP 2 can be viewed as the impact of tourism on residents and pressure on tourist facilities.

References

1. ARPA FVG 2021. I Rifiuti Urbani in Friuli Venezia Giulia - Anno (2021). <https://www.arpa.fvg.it/temi/temi/rifiuti/sezioni-principali/rifiuti-urbani/produzione-di-rifiuti-urbani-in-fvg/i-rifiuti-urbani-in-friuli-venezia-giulia-2020-primo-semester/>. Accessed July 2023
2. Defert, P.: Le taux de fonction touristique: mise au point et critique (1967)
3. Ferrari, F., Grugnale, B.: La sostenibilità del turismo nella regione Abruzzo. In: La sostenibilità del turismo nella Regione Abruzzo (2003)
4. Florence P.S.: Investment, Location and Size of Plant. University Press Cambridge, Cambridge (1948)
5. Gareth, J., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 112, p. 18. Springer, New York (2013). <https://doi.org/10.1007/978-1-4614-7138-7>
6. Giacalone, M., La Tona, L., Marino, C.: L'analisi della sostenibilità del turismo in Sicilia attraverso un indicatore globale. In: Turismo e territorio: analisi empiriche ed approcci metodologici, pp. 209–218. McGraw-Hill Education Italy SRL (2013)
7. ISTAT 2023 IStat:Turismo, <http://dati.istat.it/>. Accessed October 2023
8. ISTAT 2021, SDGs REPORT Statistical Information for 2030 Agenda in Italy. <http://dati.istat.it/>. Accessed October 2023
9. McKercher, B.: Sustainable tourism development-guiding principles for planning and management. In: National Seminar on Sustainable Tourism Development, vol. 1, No. 10 (2003)
10. REGIONE FRIULI - VENEZIA GIULIA 2023, Turismo. <https://www.regione.fvg.it/rafvfg/cms/RAFVG/economia-imprese/turismo/>. Accessed July 2023
11. REGIONE FRIULI - VENEZIA GIULIA 2022, Regione in cifre 2022. https://www.regione.fvg.it/rafvfg/export/sites/default/RAFVG/GEN/statistica/FOGLIA3/FOGLIA83/allegati/Regione_in_cifre_2022.pdf. Accessed July 2023
12. REGIONE FRIULI - VENEZIA GIULIA 2018, Decreto n° 353/AAL del 07/03/2018. https://autonomielocali.regione.fvg.it/aall/export/sites/default/AALL/news/allegati/ALLEGATI_NEWS/Decreto_n_353_del_7_marzo_2018.pdf. Accessed July 2023
13. Schoier, G., Giacalone, M.: On an analysis on some indicators of Well-Being in Italy. RIVISTA ITALIANA DI ECONOMIA, DEMOGRAFIA E STATISTICA **76**(3), 41–52 (2022)
14. WEBTUR 2023, Servizi Statistici Regione Friuli - Venezia Giulia. https://servizistatistici.regione.fvg.it/webtur/jsp/wt/login_video.jsp. Accessed July 2023



Forecasting Geneva's Donors and Their Charitable Deductions

Marta Pittavino^{1,2}(✉)  and Giedre Lideikyte-Huber³ 

¹ Venice School of Management, Ca' Foscari University of Venice, Venice, Italy
marta.pittavino@unive.it

² Geneva Centre for Philanthropy (GCP), University of Geneva, Geneva, Switzerland

³ Institute for Law and Economics, University of St. Gallen, St. Gallen, Switzerland

Abstract. The present study builds upon existing research, introducing a novel research question and innovative statistical predictions. It constitutes the first comprehensive time-series analysis of tax incentives for charitable giving among donors in Switzerland and is among the few globally to project charitable deductions. Utilizing unique panel data covering the entire Geneva taxpayers' population from 2001 to 2011, we forecast future scenarios of donors and charitable deductions. This period also coincided with a legal reform that increased ceilings for charitable deductions. Various time-series forecasting models were applied, and the model with the lowest score (indicating superior performance: best model) was chosen to project future predictions. The best model predicts a total donation amount in the range of 79 to 114 million for the years 2012–2022. These projections provide insights into the scale of deductions, aiding in decisions related to tax incentives for charitable giving. Moreover, they shed light on the key factors influencing these predictions, enabling more informed decisions on the ceiling of tax incentives for charitable contributions.

Keywords: Tax Incentives · Donors · Forecasts · ETS and ARIMA models

1 Introduction

Tax incentives for charitable giving are a prevalent feature of legal systems around the world [1]. The primary goal of such incentives, at least from an economic perspective, is to boost donations: it is for instance argued that the system that grants tax incentives for charitable donations increases transparency in the philanthropic sector [2]. However, legislative proposals are often vague on this point. Indicating that they want to increase charitable giving in general, they often fail to say how exactly donors' giving behavior is expected to change because of the legal standards establishing tax incentives for charitable giving or which donors these reforms intend to benefit [3].

To date, it is still not entirely known how well tax incentives for charitable donations succeed in influencing taxpayers' behavior. Relatively few studies, particularly in Europe, examine the characteristics of donors who respond to tax incentives for charitable giving, and even fewer studies concentrate on changes in donors' reactions to actual changes in tax law.

Current Swiss law allows taxpayers deducting, up to a certain threshold, charitable contributions from their taxable income for individuals and for corporations. This deduction is subject to a threshold which is currently 20% of the net taxable income or profits for federal income tax purpose, with a minimum donation requirement of 100 CHF [4]. According to the Federal Constitution, cantons can fix their own income tax rates, but most of the cantonal tax thresholds are also fixed at same threshold of 20% [5]. The 20% federal deduction threshold was introduced on 1 January 2006 as part of a larger reform of the Swiss federal law [6], replacing the previous threshold of 10%. This reform has also carried out other major modifications of federal tax law norms related to charitable giving, introducing a deduction of charitable non-cash donations was introduced and allowing the deductions of donations to the Swiss Confederation, the cantons, the communes, and their institutions [6]. After the reform at the federal level, cantonal law modifications were introduced; in the canton of Geneva, the 5% deduction threshold of taxable net individual income [7] increased to 20% in 2010 [7] and the deduction threshold for corporations increased from 10% to 20% [8].

In the work published in [9], the authors of this paper analyzed the characteristics of donors' giving in relation to their income and in the framework of an income tax law reform. The objective of the present paper shifts focus from the legal context to the time influence by studying the timing of giving, concentrating on future projections of charitable deductions amounts and a number of donors, based on previous trends.

2 Material and Methods

2.1 Data Description

Our analysis is based on information from taxpayer returns over 11 years from the year 2001 until 2011 that the Tax Administration of the Canton of Geneva (TACG) confidentially gave us for this study. The selected variables provide information on the entire population of taxpayers in the Canton of Geneva (approximately 250'000 households). A different data set was provided for each year under study, 11 in total. Each data set comprised the same variables, an entire description of them is provided in [3, 9] the specific ones particularly used in the present study are described and listed below with their original name provided in brackets. For this specific follow-up study, a new variable, "year", has been specifically created to enable a more in-depth longitudinal analysis for the current and future year of study, based on the panel data forecasting literature [10, 11]. A merging of the 11 different data sets, with the elimination of double IDs, if any, was performed to create an appropriate unique dataset:

- "*coded ID*" ("*identifiant*"): a coded ID for each taxpayers. This variable allows to follow the same taxpayer over time. The same coded ID is used for a given taxpayer for each fiscal year. As Switzerland has a joint filing system, married couples are considered and treated as one taxpayer in the same way as a single non-married individual, and they have only one coded ID. In this paper, any deducting taxpayer, couple, or individual is referred to as "*donor*" and it represents a subset of "*coded ID*", taxpayers' population we investigate.

- “*deductions for donations*” (“*versements_benevoles*”): the amount of deduction for charitable giving, representing the entire annual amount of the deducted donations or capped amount, if exceeding the deductible threshold.
- “*year under study*” (“*year*”): this new variable has been generated for the purpose of this study to keep track of the evolution by year of the collected data. It indicates the 11 years for this study, from year 2001 to year 2011.

This data was selected for taxpayers residing in the Canton of Geneva as well as for taxpayers residing in another Swiss canton or abroad, however still taxed in Geneva due to their limited taxing liability there.

As reported in the previous study [3], the total number of taxpayers in the canton of Geneva has steadily increased, from 234,117 in 2001 to 266,336 in 2011. The share of the taxpayers deducting charitable donations more than doubled, passing from 8.3% in 2001 to 19.3% in 2011, with a steep increase in 2005 (deducting taxpayers reaching 16.3%). Concerning the general pattern of deductions during the studied period, the total amount of yearly charitable deductions increased significantly, from CHF 29,133,697 in 2001 to CHF 72,741,235 in 2011 which is due to the rise in population and a substantial increase of 48% is recorded in 2009.

2.2 Methods

Forecasting methods applied to the amounts of deductions and the number of donors, to project the predicted figures in the next upcoming 10 years: 2012–2021.

The class of methods used to make predications of the current values for the next upcoming 10 years were the ETS (Error-Trend-Seasonality) and the ARIMA Models [12]. Four different types of ETS models (from Simple Exponential Smoothing (SES) to Holt's Models), considering different trend effects (i.e. additive, additive with damped and multiplicative) since the presence of seasonality was not detected, have been fitted to compute the amount of donations for the upcoming 10 years (2012–2022). Five different error metrics (i.e. AIC, AICc, BIC, RMSE and MAPE) have been calculated and the smallest ones have been implied to select the best model for the forecast.

3 Results

Five forecasting methods: ETS model with several trend effects, have been implemented and compared using the error metrics to identify the one that is the most suitable to predict the amounts of deductions and with the best data fit. ARIMA models have also been fitted, however they behaved as SES by always producing the same output, therefore they have been discarded for the donations' forecast. While they have been incorporated for donors' forecast.

The model which performed the best, with the lowest AIC, BIC and MAPE error metrics term, was model 4) in Table 1 related to a Holt's model with additive damped effect and a multiplicative error to incorporate the increasing nature of the donations over time. The result of this forecast over the predicted period can be seen in Fig. 1, where the previously mentioned model with the 80% and 95% prediction intervals is shown (Table 2).

Table 1. Comparison of the 5 forecasting models (simple exponential smoothing, Holt’s model with and without damped effect for the trend) implemented for donations’ projections and their error metrics.

Forecasting models	Errors				
	<i>AIC</i>	<i>AICc</i>	<i>BIC</i>	<i>RMSE</i>	<i>MAPE</i>
1) <i>SES</i>	389.1287	392.5573	390.3224	11’028’513	16.50398
2) <i>ETS(A, A, N)</i>	381.3582	393.3582	383.3477	6’458’918	10.24257
3) <i>ETS(A, Ad, N)</i>	383.3466	404.3466	385.7340	6’455’514	9.902257
4) <i>ETS(Z, Ad, N)</i>	378.9578	399.9578	381.3451	6’511’752	9.147868
5) <i>ETS(M, M, N)</i>	383.5166	395.5166	385.5061	8’064’655	12.53688

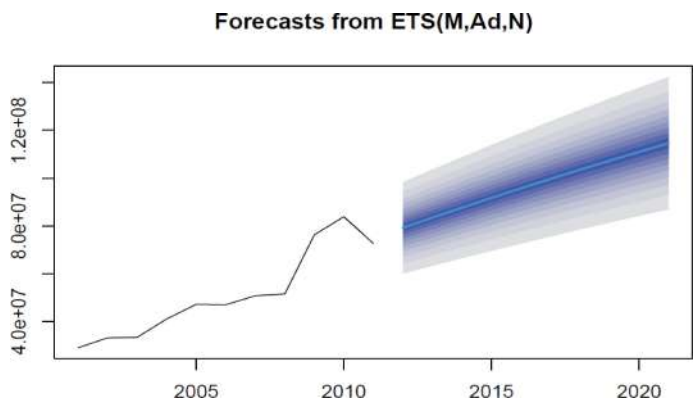


Fig. 1. Donations’ forecasts, over the year 2012–2021. The donations’ forecast trend is moving from 79’170’591 CHF in 2012 to 114’651’701 CHF in 2021.

Table 2. Projections of the amount of donations over the following 10 years (2012–2021), resulting from the model ETS(M, Ad, N) with 80% and 95% prediction intervals.

Years	Forecast	Lo 80% PI	Hi 80% PI	Lo 95% PI	Hi 95% PI
2012	79’170’591	64’241’346	94’099’835	56’338’282	102’002’900
2013	83’438’939	67’704’809	99’173’068	59’375’664	107’502’213
2014	87’621’919	71’099’002	104’144’837	62’352’298	112’891’541
2015	91’721’240	74’425’310	109’017’170	65’269’398	118’173’082
2016	95’738’574	77’685’091	113’792’057	68’128’155	123’348’994
2017	99’675’561	80’879’676	118’471’447	70’929’736	128’421’387

(continued)

Table 2. (continued)

Years	Forecast	Lo 80% PI	Hi 80% PI	Lo 95% PI	Hi 95% PI
2018	103'533'809	84'010'368	123'057'249	73'675'283	133'392'334
2019	107'314'891	87'078'445	127'551'336	76'365'918	138'263'863
2020	111'020'351	90'085'160	131'955'542	79'002'739	143'037'962
2021	114'651'701	93'031'739	136'271'664	81'586'822	147'716'581

We did not forecast more than 10 years, because since we had information from the previous 11 years, we did not want to go too far in time to limit the uncertainty (Fig. 2 and Table 3).

Table 3. Comparison of the 5 forecasting models (simple exponential smoothing and Holt's model with and without damped effect for the trend, several errors and ARIMA model) implemented for donors' projections and their error metrics.

Forecasting models	Errors				
	<i>AIC</i>	<i>AICc</i>	<i>BIC</i>	<i>RMSE</i>	<i>MAPE</i>
1) <i>SES</i>	211.6988	215.1274	212.8925	3466.838	8.264631
2) <i>ETS(A,A,N)</i>	203.8104	215.8104	205.7999	2019.522	5.63453
3) <i>ETS(A,Ad,N)</i>	204.6256	225.6256	207.0130	1913.642	5.161422
4) <i>ETS(Z,Ad,N)</i>	201.7138	213.7138	203.7033	1963.346	3.66038
5) <i>ARIMA(0,2,0)</i>	162.16	162.73	162.36	1601.337	2.202275

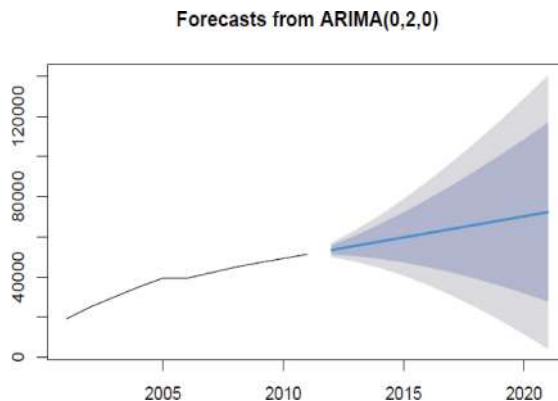


Fig. 2. Donors' forecasts, over the year 2012–2021. The donors' forecast trend is ranging from 53'485 donors in 2012 to 72'349 donors in 2021.

For the donors' forecast, given the data pattern, a multiplicative error for a model with additive damped trend would not be supported, giving similar results as model 3)

with an additive error instead. The best model to forecast the donors' prediction was an ARIMA model double differentiated, with no autoregressive or moving average part (Table 4).

Table 4. Projections of the amount of donors over the following 10 years (2012–2021), resulting from the model ARIMA(0, 2, 0) with 80% and 95% prediction intervals.

Years	Forecasts	Lo 80% PI	Hi 80% PI	Lo 95% PI	Hi 95% PI
2012	53485	51216.21	55753.79	50015.19	56954.81
2013	55581	50507.84	60654.16	47822.27	63339.73
2014	57677	49187.97	66166.03	44694.15	70659.85
2015	59773	47346.34	72199.66	40768.06	78777.94
2016	61869	45043.22	78694.78	36136.19	87601.81
2017	63965	42322.14	85607.86	30865.11	97064.89
2018	66061	39216.34	92905.66	25005.64	107116.36
2019	68157	35752.23	100561.77	18598.18	117715.82
2020	70253	31951.45	108554.55	11675.84	128830.16
2021	72349	27832.17	116865.83	4266.38	140431.62

4 Discussion and Conclusion

This study builds up on [9] by focusing on forecasts future charitable deductions and donors that use tax incentives for charitable donations.

In terms of the future deductions, we estimate an increase in the deductions of charitable donations in the canton of Geneva, reaching around CHF 114'651'701 in 2021 (57,6% increase from 2011). It must be highlighted that the only estimates we could find are the ones for the total amount of private donations, and not only charitable donations, by the Swiss population which was more than CHF 2.05 billion in 2021 [13]. In 2021, the Swiss population was 8.7 million and Geneva had 511'921 inhabitants (not taxpayers). This represents 6% of the total Swiss population, 6% of the 2.05 billion is CHF 123 million: outcome slightly higher than our forecast, but comprehensive of total donations and not only deductions. Thus, ours are close and unique estimate.

In terms of future donors, we estimate that the donors will continue to increase in an important manner in Geneva, reaching 72'349 donors in 2021, representing almost four times the initial number of donors in 2001 (19'335).

References

1. OCDE. Taxation and Philanthropy, OECD Tax Policy Studies, n° 27, Éditions OCDE, Paris (2020). <https://doi.org/10.1787/df434a77-en>
2. Brakman, R.D., Dean S.A.: For-Profit Philanthropy: Elite Power and the Threat of Limited Liability Companies, Donor-Advised Funds, and Strategic Corporate Giving. Oxford University Press, Oxford (2023)
3. Lideikyte-Huber, G., Pittavino, M., Peter, H.: Tax incentives for charitable giving: 2006 Reform and Geneva based evidence. In: Peter, H., Lideikyte Huber, G. (eds.) Routledge Handbook of Taxation and Philanthropy (2021)
4. Federal Act of 14 December 1990 on Direct Federal Taxation (DFTA), AS 1991 1184
5. The Federal Constitution of the Swiss Confederation of 18 April 1999, AS 1999 2556
6. Code civil suisse (Droit des fondations), Modification du 8 octobre 2004, RS 4545 (Swiss civil code - modifications)
7. Personal Income Tax Act - V (FR : Loi sur l'imposition des personnes physiques (LIPP-V) Détermination du revenu net – Calcul de l'impôt et rabais d'impôt - Compensation des effets de la progression à froid du 22 septembre 2000, D3 16 - D3 08; in force: 01.01.2001
8. Corporate Income Tax Act of September 23, 1994 (FR : Loi sur l'imposition des personnes morales du 23 septembre 1994, D 3 15; in force: 01.01.1995)
9. Lideikyte-Huber, G., Pittavino, M.: Who donates and how? New evidence on the tax incentives in the canton of Geneva, Switzerland. *J. Empir. Leg. Stud.* **19**(3), 758–797 (2022)
10. Baltagi, B.H., Pesaran, M.H.: Heterogeneity and cross section dependence in panel data models: theory and applications. *J. Appl. Econ.* **22**, 229–232 (2007)
11. Baltagi, B.H.: *Econometric Analysis of Panel Data*. Springer, Cham (2021). <https://doi.org/10.1007/978-3-030-53953-5>
12. Hyndman, R.J., Athanopoulos, G.: *Forecasting: Principles and Practice*, 3rd edn. OTexts, Melbourne (2021)
13. Spendenreport Schweiz (2022)



Bayesian Meta-learning Approach for Feasible Large Spatial Analysis

Luca Presicce¹(✉) and Sudipto Banerjee²

¹ Department of Management, Economics and Statistics, University of
Milano-Bicocca, Milan, Italy

l.presicce@campus.unimib.it

² Department of Biostatistics, University of California Los Angeles,
Los Angeles, CA, USA

Abstract. Geostatistical modeling is afflicted by onerous computational effort when the number of observed locations is very large. While there exists a burgeoning literature today that attempts to tackle the so-called “big-n” problems, spatial inference remains unfeasible for moderate data sets on modest computing architectures. Our current contribution resides in the domain of “meta-” approaches where a massive data set is split into smaller data sets, each data set is analyzed independently and the inference from these individual data are combined to approximate fully model-based Bayesian inference. Our specific contribution is to introduce Bayesian predictive stacking in spatial meta-analysis in the context of univariate spatial data. Furthermore, we aim to make inference and uncertainty quantification feasible without excessively demanding hardware settings. Introducing new methodologies, and exploiting existing techniques, the analysis of a massive data set with observations in millions is illustrated.

Keywords: Bayesian predictive stacking · meta-learning · geostatistical modeling · conjugate models

1 Introduction

A primary challenge in spatial data science is the analysis of big georeferenced sets of data. This stems from the demanding Gaussian likelihood computations involving matrix factorizations and determinant computations for large spatial covariance matrices. This is referred to as the “Big N” problem in spatial statistics. In the same way, spatial prediction, also called kriging, relies on Gaussian process regression methodologies, that scale very badly with massive sets of data. To manage this problem, burgeoning literature on the analysis of large spatial data sets focuses mostly on a few bins of solutions, at whose foundation there are common ideas. Many of them are in the direction of some sort of dimensionality reduction. One customary idea aims to reduce the dimension of the spatial covariance matrix by imposing a low-rank structure. A second class

of approaches considers sparsity as a solution as well. However, they assume sparsity over the inverse of the spatial covariance matrix, usually achieved by imposing conditional independence, or also with composite likelihood. Another school of thought works reduction by considering smaller problems instead. Partitioning the domain into subregions, and then hierarchically combining models ensures the borrowing of information. Even if there are similarities with the last, the current contribution follows a different idea. The here-introduced accelerated spatial meta-kriging (ASMK) lies in a class of methodology that more recently appeared in the literature, as [4, 5]. These contributions, chase a meta-learning-based approach, exploiting Divide-and-Conquer procedures.

The contents of the paper evolves as follows. In Sect. 2 we briefly present the proposed accelerated spatial meta-kriging approach. Section 3, illustrate an analysis on sea surface temperature with over 1 million observed locations. Conclusive considerations end the work in Sect. 4.

2 Accelerating Spatial Meta-Kriging

To better understand the contents in Sect. 2, we remind the reader to the concepts of Bayesian predictive stacking (BPS) of predictive densities. A comprehensive dissertation can be found in [6], and for further details see also [3]. Hereafter, we introduce the accelerated spatial meta-kriging approach, for univariate modeling. The ASMK is a Divide-Conquer strategy, composed as follows. In the divide step, after partitioning the data, we perform the BPS on each subset (within-subsets), obtaining a set of stacked posterior distributions. Then in the conquer step, we serve of BPS procedure, but on stacked posterior distributions (between-subsets), to combine them. By first BPS, we obtain subset inferences, while the second BPS provides full posterior inference. Let us consider a customary regression model for a spatially indexed outcome $y(s)$ at a location s in a bounded region $\mathcal{D} \subset \mathbb{R}^d$,

$$y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \beta + \omega(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad (1)$$

where $\mathbf{x}(\mathbf{s})$ is a $p \times 1$ vector of spatially referenced predictors, β is a $p \times 1$ vector of slopes measuring the trend, $\omega(\mathbf{s}) \sim \text{GP}(0, \sigma^2 \boldsymbol{\rho}_\psi(\cdot, \cdot))$ is a zero-centred spatial Gaussian process on $\mathbb{R}^d$ with spatial correlation function $\boldsymbol{\rho}_\psi(\cdot, \cdot)$ depending on spatial range parameter ψ , and σ^2 is a scale (spatial variance) parameter. The white noise process $\varepsilon(\mathbf{s}) \sim \text{N}(0, \tau^2)$ with variance τ^2 captures measurement error. Let $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \in \mathcal{D}$ be a set of n spatial locations, yielding to $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))$ with predictors at these locations collected in the $n \times p$ matrix $\mathbf{X} = [\mathbf{x}(\mathbf{s}_1) : \dots : \mathbf{x}(\mathbf{s}_n)]^\top$. Defined the finite-dimensional realization of the spatial process as $\boldsymbol{\omega} = (\omega(\mathbf{s}_1), \dots, \omega(\mathbf{s}_n))^\top$, and let $\boldsymbol{\rho}_\psi(\mathcal{S}, \mathcal{S})$ be the $n \times n$ spatial correlation matrix constructed from $\boldsymbol{\rho}_\psi(\cdot, \cdot)$. A Bayesian hierarchical model, including the specification of prior distributions, is constructed as

$$\mathbf{y} \mid \boldsymbol{\omega}, \beta, \sigma^2 \sim N(\mathbf{X}\beta + \boldsymbol{\omega}, \delta^2 \sigma^2 \mathbf{1}_n), \quad (2)$$

$$\boldsymbol{\omega} \mid \sigma^2 \sim N(0, \sigma^2 \boldsymbol{\rho}_\psi(\mathcal{S}, \mathcal{S})) \quad (3)$$

$$\beta \mid \sigma^2 \sim N(\mu_\beta, \sigma^2 V_\beta), \quad \sigma^2 \sim \text{IG}(a_\sigma, b_\sigma). \quad (4)$$

where the spatial correlation parameter ψ and the noise-to-spatial variance ratio $\delta^2 := \tau^2/\sigma^2$ are considered fixed. While $\mu_\beta, V_\beta, a_\sigma$, and b_σ are fixed hyperparameters specifying the prior distributions for β and σ^2 . Considering to fix $\{\psi, \delta^2\}$, ensures closed-form conjugate posterior inference for this hierarchical specification, for further details on the results presented below see [1, 2].

Let $\boldsymbol{\gamma} = [\beta^\top, \boldsymbol{\omega}^\top]^\top$, then a conjugate Bayesian model is obtained by considering the joint prior on $\{\boldsymbol{\gamma}, \sigma^2\}$, denoted as $p(\boldsymbol{\gamma}, \sigma^2 \mid \mu_\beta, V_\beta, a_\sigma, b_\sigma)$. For any fixed $\{\psi, \delta^2\}$, we have the closed-form posterior density

$$p(\boldsymbol{\gamma}, \sigma^2 \mid \mathbf{y}) = p(\sigma^2 \mid \mathbf{y}) p(\boldsymbol{\gamma} \mid \sigma^2, \mathbf{y}) \quad (5)$$

$$= \text{IG}(\sigma^2 \mid a_\sigma^*, b_\sigma^*) N(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \sigma^2 M_\star), \quad (6)$$

$$\text{where } a_\sigma^* = a_\sigma + n/2, \quad b_\sigma^* = b_\sigma + 1/2 (\mathbf{y}_\star - \mathbf{X}_\star \hat{\boldsymbol{\gamma}})^\top V_\star^{-1} (\mathbf{y}_\star - \mathbf{X}_\star \hat{\boldsymbol{\gamma}}), \quad (7)$$

$$M_\star^{-1} = \mathbf{X}_\star^\top V_\star^{-1} \mathbf{X}_\star, \quad \hat{\boldsymbol{\gamma}} = M_\star \mathbf{X}_\star^\top V_\star^{-1} \mathbf{y}_\star, \quad (8)$$

where $\mathbf{y}_\star, \mathbf{X}_\star, V_\star$ composed the augmented linear system cast from the spatial model in Eq. (2). For more specifications about these results see [1, 9]. Let $\mathcal{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_{n'}\} \in \mathcal{D}$ be a set of n' of unknown points, with $\boldsymbol{\omega}_\mathcal{U}$ and $\mathbf{y}_\mathcal{U}$ be the $n' \times 1$ vectors with elements $\omega(\mathbf{u}_i)$ and $y(\mathbf{u}_i)$ for $i = 1, 2, \dots, n'$. Let $\mathbf{X}_\mathcal{U} = (\mathbf{x}(\mathbf{u}_1) : \dots : \mathbf{x}(\mathbf{u}_{n'}))^\top$ be the $n' \times p$ matrix of predictors and let $\boldsymbol{\rho}_\psi(\mathcal{U}, \mathcal{U})$ the spatial correlation matrix at $\mathcal{U}$. Then, Bayesian inference proceeds from exact posterior samples obtained from Eq. (5). We first draw values of $\sigma^2 \sim \text{IG}(a_\sigma^*, b_\sigma^*)$ followed by a single draw of $\boldsymbol{\gamma} \sim N(\hat{\boldsymbol{\gamma}}, \sigma^2 M_\star)$ for each drawn value of σ^2 . This yields samples $\{\boldsymbol{\gamma}, \sigma^2\}$ from Eq. (5). While predictive inference for the latent process $\boldsymbol{\omega}_\mathcal{U}$ and the outcome $\mathbf{y}_\mathcal{U}$ is obtained by drawing a value of $\boldsymbol{\omega}_\mathcal{U} \sim N(\mu_\omega(\boldsymbol{\gamma}), \sigma^2 V_\omega)$ with $\mu_\omega(\boldsymbol{\gamma}) := \boldsymbol{\rho}_\psi(\mathcal{U}, \mathcal{S}) \boldsymbol{\rho}_\psi(\mathcal{S}, \mathcal{S})^{-1} \boldsymbol{\omega}$ and $V_\omega := \boldsymbol{\rho}_\psi(\mathcal{U}, \mathcal{U}) - \boldsymbol{\rho}_\psi(\mathcal{U}, \mathcal{S}) \boldsymbol{\rho}_\psi(\mathcal{S}, \mathcal{S})^{-1} \boldsymbol{\rho}_\psi^\top(\mathcal{U}, \mathcal{S})$ for each value of $\{\boldsymbol{\gamma}, \sigma^2\}$ drawn above (see Sect. 3.4 in [1]), then drawing a value of $\mathbf{y}_\mathcal{U} \sim N(\mathbf{X}_\mathcal{U} \beta + \boldsymbol{\omega}_\mathcal{U}, \sigma^2 \delta^2 \mathbf{1}_{n'})$ for each drawn value of β (extracted from $\boldsymbol{\gamma}$), σ^2 and $\boldsymbol{\omega}_\mathcal{U}$. As anticipated, this tractability is only possible if the range decay ψ and the noise-to-spatial variance ratio δ^2 are fixed. So far, alternative approaches using K -fold cross-validation have been explored with limited success [2]. The aforementioned conjugate modeling is then used in each subset of the data. Given this framework, in the Divide-step we use BPS to “accelerate” the subset modeling, and then the entire inference. Indeed, BPS is a particular case of Bayesian model averaging approaches, which allows us to “integrate out” the hyperparameters $\{\delta^2, \psi\}$. This avoids the need for simulation-based approaches, such as MCMC, exploiting exact conjugate modeling, and resulting in an acceleration of the inferences. However, as presented in [6], to perform BPS of predictive densities one has to solve the optimization problem:

$$\max_{\mathbf{z}_k \in \mathbf{S}_1^J} \frac{1}{n_k} \sum_{i=1}^{n_k} \log \sum_{j=1}^J z_{k,j} p(y_{k,i} | \mathbf{y}_{k,-i}, \mathbf{M}_j), \quad (9)$$

where $\mathbf{S}_1^J$ is the J -dimensional simplex, and J represents the number of model configurations $\mathbf{M}_j$, i.e. the cardinality of the set composed by a grid of hyperparameter $\{\delta^2, \psi\}$ values. In Equation (9), the quantities $\mathbf{y}_k$ represent the subvector of $\mathbf{y}$ that corresponds to the k -th subset, n_k is its length, and $y_{k,i}, \mathbf{y}_{k,-i}$ are the i -th element of $\mathbf{y}_k$ and the vector itself without the i -th element, respectively. Once the stacking weights $\hat{\mathbf{z}}_k = \{\hat{z}_{k,j}\}_{j=1,\dots,J}$, for $k = 1, \dots, K$ are obtained, the stacked estimate of the predictive density, within each partition, can be recovered as

$$\hat{p}(\tilde{\mathbf{y}} | \mathbf{y}_k) = \sum_{j=1}^J \hat{z}_{k,j} p(\tilde{\mathbf{y}} | \mathbf{y}_k, \mathbf{M}_j), \quad k = 1, \dots, K, \quad j = 1, \dots, J. \quad (10)$$

Thus, for each of the K subsets, the BPS procedure provides an estimate of the posterior predictive $\hat{p}(\tilde{\mathbf{y}} | \mathbf{y}_k)$, and a set of stacking weights $\hat{\mathbf{z}}_k$. To recover the global inference for the entire data set, we apply the BPS procedure again. Hence, a double stacking will be performed. Then, the double BPS of predictive densities boils down to the solution of the following convex optimization problem:

$$\max_{\mathbf{w} \in \mathbf{S}_1^K} \frac{1}{n} \sum_{i=1}^n \log \sum_{k=1}^K w_k \sum_{j=1}^J \hat{z}_{k,j} p(y_{k,i} | \mathbf{y}_{k,-i}, \mathbf{M}_j), \quad (11)$$

with $\mathbf{S}_1^K$ as the K -dimensional simplex. Again, once the stacking weights $\mathbf{w} = \{w_k\}_{k=1,\dots,K}$, that regulate the combination of the individual subset posterior predictive models are obtained by Equation (11), the double BPS estimation of the full posterior predictive distribution is then recovered:

$$\hat{p}(\tilde{\mathbf{y}} | \mathbf{y}) = \sum_{k=1}^K \hat{w}_k \hat{p}(\tilde{\mathbf{y}} | \mathbf{y}_k), \quad k = 1, \dots, K. \quad (12)$$

Let $\mathcal{U}$ be a set of n' of unknown points as in Sect. 2. Thus, the prediction for the new response $\mathbf{y}_{\mathcal{U}}$ can be straightforwardly attained from Equation (12).

$$\hat{p}(\mathbf{y}_{\mathcal{U}} | \mathbf{y}) = \sum_{k=1}^K \hat{w}_k \sum_{j=1}^J \hat{z}_{k,j} p(\mathbf{y}_{\mathcal{U}} | \mathbf{y}_k, \mathbf{M}_j). \quad (13)$$

A similar strategy can be adopted for the posterior predictive surface for the spatial process $\boldsymbol{\omega}$. Indeed, the prediction for $\boldsymbol{\omega}_{\mathcal{U}}$ is derived from

$$\hat{p}(\boldsymbol{\omega}_{\mathcal{U}} | \mathbf{y}) = \sum_{k=1}^K \hat{w}_k \hat{p}(\boldsymbol{\omega}_{\mathcal{U}} | \mathbf{y}_k), \quad (14)$$

where $\hat{p}(\boldsymbol{\omega}_{\mathcal{U}} | \mathbf{y}_k) = \sum_{j=1}^J \hat{z}_{k,j} p(\boldsymbol{\omega}_{\mathcal{U}} | \mathbf{y}_k, \mathbf{M}_j)$, and $\{\hat{z}_{k,j}\}$ as the same stacking weights obtained before, for $k = 1, \dots, K$ and $j = 1, \dots, J$.

3 Real Data Illustration

One of the most discussed topics of the last years is global warming and its consequences. Then, institutions have started to monitor key aspects of its evolution, providing data that allows the research to develop crucial predictive global models. One of the main aspects is the surface sea temperature (SST), which NASA maintains a satellite-based database that extends all over the world. Further details on this can be found on [NASA](#) related online portal. In this section, we present an application of ASMK to the SST data. The current offer has the scope to enable researchers to perform geostatistical analysis on massive spatial data sets, even with modest computational and memory resources. For this reason, the data analysis was carried out on a standard laptop. In detail, we used a Windows machine equipped with 16 Gb of RAM, Intel i7-8750H processor (enabling parallel computations over 10 logical cores). As an illustration, we analyze a data set composed of 1,001,000 georeferenced observations of SST, collected within June 2017. To this end, we take advantage of the data set used in [7], but differently from them, we consider SST data all over the world, instead of a selected subregion. Within this massive data set, we use $n = 1,000,000$ observations for model fitting while the rest for predictive assessment. As explanatory variables, we consider the coordinates after a sinusoidal projection. For how to concern the Divide-Conquer strategy, we take into account $K = 2,000$ subsets of random sampled observations, that corresponds to a subset size of 500 units. Prior choices follow standard quantities, as considered in [7], and [9]. The set of hyperparameter values was defined by considering explanatory spatial data analysis, evaluating the variograms to find insight about ψ and δ^2 . In Fig. 1, the interpolated maps of the results are reported.

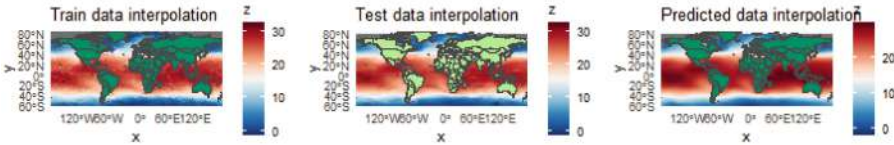


Fig. 1. From left to right: comparison between training, test, and predicted surface.

Figure 1 shows an empirically striking capability to reproduce the test data surface of SST. Moreover, the full model-based Bayesian inference requires only 280 min, just more than 4 h. We assess predictive performance using RMSPE on 1000 withheld observations for the ASMK model, comparing it with the Bayesian linear model (BLM), which does not account for spatial variability. While the BLM performs a RMSPE of 8.981, the accelerated spatial meta-kriging approach achieves a much lower RMSPE of 1.346. This remarks the effectiveness of the proposed procedure, as well as the chance to have Bayesian geostatistical modeling feasible with millions of observations. The entire analysis and graphics were

implemented in native R and c++ programming languages within the Rcpp package, indeed called ASMK, freely available on the corresponding author GitHub repository [lucapresicce/ASMK](https://github.com/lucapresicce/ASMK).

4 On Going Features

Here, we briefly illustrate the principles behind the ASMK for univariate geostatistical modeling. This is currently an ongoing work, and as such, some features need more investigation and elaboration. First of all, a set of simulations are under development now. In particular, we empirically want to show that ASMK is capable of recovering spatial surfaces, without penalizing the posterior learning of the parameters, mainly checking for the inference about the regression coefficients β and the variance σ^2 . Moreover, we aim to show that ASMK behaves as the SMK of [4], but that it is faster. Indeed, the reduced computational complexity and the better timing are two of the main strengths of ASMK. Furthermore, ASMK pretends to make feasible the analysis of massive spatial data sets even on modest computational resources. To this matter, the parallelization framework is self-constrained to a standard laptop, as explained in Sect. 3. The motivation can be traced back to the fact that most of the solutions in the literature use huge computational infrastructures. However, this does not allow computational advantages reproducible in general. As evident from the illustration in Sect. 3, we could be on the right path. Indeed, we provide a feasible approach to the analysis of millions of points, in a reasonable amount of time without compromising the inferences, and without using extreme computational resources.

References

1. Banerjee, S.: Modeling massive spatial datasets using conjugate Bayesian linear modeling framework. *Spat. Stat.* **37** (2020). <https://doi.org/10.1016/j.spasta.2020.100417>
2. Finley, A.O., Datta, A., Cook, B.D., Morton, D.C., Andersen, H.E., Banerjee, S.: Efficient algorithms for bayesian nearest neighbor gaussian processes. *J. Comput. Graph. Stat.* **28** (2019). <https://doi.org/10.1080/10618600.2018.1537924>
3. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** (2007). <https://doi.org/10.1198/016214506000001437>
4. Guhaniyogi, R., Banerjee, S.: Meta-kriging: scalable Bayesian modeling and inference for massive spatial datasets. *Technometrics* **60** (2018). <https://doi.org/10.1080/00401706.2018.1437474>
5. Guhaniyogi, R., Li, C., Savitsky, T., Srivastava, S.: A divide-and-conquer Bayesian approach to large-scale kriging. Submitted to *Stat. Sci.* (2019). <https://doi.org/10.48550/arXiv.1712.09767>
6. Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Anal.* **13** (2018). <https://doi.org/10.1214/17-BA1091>
7. Zhang, L., Datta, A., Banerjee, S.: Practical Bayesian modeling and inference for massive spatial data sets on modest computing environments. *Stat. Anal. Data Min.* **12** (2019). <https://doi.org/10.1002/sam.11413>

8. Zhang, L., Banerjee, S.: Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data. *Biometrics* **78** (2021). <https://doi.org/10.1111/biom.13452>
9. Zhang, L., Tang, W., Banerjee, S.: Exact bayesian geostatistics using predictive stacking. arXiv preprint. arxiv.org/abs/2304.12414 (2023)



Social Inequality and Student Mobility from Southern Italian Regions

Andrea Priulla¹(✉), Eleonora Miaci², and Nazareno Panichella²

¹ Department of Economics, Business, and Statistics, University of Palermo,
Palermo, Italy

andrea.priulla@unipa.it

² Department of Social and Political Sciences, University of Milan, Milan, Italy

Abstract. The North-South disparities in terms of economic development in Italy are also reflected in the inequalities in educational opportunities. Southern regions, already plagued by significant school dropout rates, experience significant mobility student outflows at the university level. In this context, this work aims to study the interplay of sociodemographic factors and others related to previous educational attainment on the mobility choices of southern graduates. In particular, the study focuses on two possible choices: enrollment in a southern university or a central-northern one. The results highlight how mobility choices are strongly differentiated based on the interaction of social class and the student's previous educational attainment.

Keywords: social background · high school track · academic ability · student mobility

1 Introduction

Educational outcomes are affected by an intricate net of factors, among which socioeconomic background plays a fundamental role. Despite the detection of a decline of inequality in educational opportunities in several countries [1], the family background still represents an important predicting factor for individual attainments [2]. Research conducted in Italy has revealed that the decision to engage in higher education and subsequent mobility decisions are strongly tied to students' socioeconomic backgrounds, in addition to their abilities and personal characteristics. Additionally, inequalities in educational opportunities are reflected in the North-South divide characterising many facets of our country. A recent report by the Bank of Italy [3] has indeed underlined the lower propensity of southern graduates to continue their studies after graduation compared to other Italian graduates. At the same time, the AlmaLaurea Report (2023) highlights that about 28% of southern students opt for universities located in the Centre-North. This mobility choice is often driven by the greater prestige and better employment prospects associated with a degree from those universities. Research has indeed shown that territorial mobility significantly improves

the job market insertion for individuals from the South: students who move to northern regions to complete their studies (*ante-lauream*) register higher employment rates and better earnings compared to their peers who did not undertake such mobility [4]. Impicciatore and Tosi [5] have pointed out that mobility from the South to the North is further influenced by contextual elements such as the social class and family backgrounds of the students. In this process, families are often required to provide financial support to their children, viewing it as an investment in their future social progress. The mobility choice involves significant financial obstacles for southern students, especially due to the higher costs of living, particularly housing costs, in the central and northern areas [6]. Furthermore, the migration is generally selective of students with better academic performance, resulting in further impoverishment of the human capital of the South [3, 7]. Starting from the premise, the aim is to study the enrolment choices of southern graduates according to two potential choices: the choice to move to central northern regions or to remain in the same region where they graduated from high school. In detail, we delve into the analysis of the interplay between social origin, academic ability, and educational track on mobility choices.

2 Theoretical Framework

Several international studies revealed how access to better educational opportunities is stratified by social origins [8], because of the constraints to mobility associated with a disadvantaged family background [5]. It has been shown that social origin disparities differently influence educational transitions, becoming more pronounced when there is a greater risk of downward social mobility and loss of income. While the impact of social origin on educational paths has received considerable attention, the role of academic ability and its interplay with social background has been less thoroughly examined. Bernardi [9], in the Spanish context, investigates the compensatory effect of social background in correcting educational missteps and examines how cognitive and non-cognitive skills can influence the effect of social background on educational transitions. Family resources, particularly in higher social classes, can increase the likelihood of educational success, whereas, only the most talented students from disadvantaged backgrounds manage to reach a higher educational level. The choice of high school track plays a crucial role in shaping students' academic careers, especially in a stratified educational system such as the Italian one. This choice is strongly associated with family social background since the family is usually responsible for their children's education at this stage. For instance, [10] demonstrated that students attending lyceums are much more likely to leave southern regions. Moreover, track selection inevitably determines students' field of study choice in tertiary education. This provides a further boost to mobility due to the uneven regional distribution of degree programs in Italy. In this framework, the regional variation in the value of educational credentials can drive students to relocate to areas where their credentials are more highly valued or better aligned with local job market demands. The interplay between social origin, academic

ability, and educational track is pivotal in the literature on social stratification. There is instead a paucity of research focusing on how this interplay influences geographical mobility [11].

3 Research Questions and Hypotheses

Based on the stratification literature, we have hypothesized two possible scenarios: university study in the North as a “boosting” or as a “compensatory advantage”. In the first scenario, families invest in already well-prepared children, aiming to significantly elevate their chances of success. In the second scenario, investments are made for the child who requires more validation. This is because the most prepared children can establish themselves in the job market even after studying at a less prestigious university in the South.

Below are the research questions and related hypotheses we have formulated:

- Q1: Are Southern students from better-off families and with high academic ability more likely to enrol in a central-northern university?
- H1: We expect a positive association between social origin and mobility. Given the lack of literature on the topic, we do not have expectations regarding the role of academic ability.
- Q2: Does the effect of academic ability vary among Southerners with different social origins?
- H2: We hypothesize that the propensity to move is higher among students with better academic ability (“boosting effect”).

4 Data

The analysis is based on the linkage of micro-data coming from two national administrative archives: INVALSI and the *Anagrafe Nazionale Studenti* (ANS) [12]. INVALSI carries out annual national standardized tests at five educational levels to evaluate student ability in mathematics, Italian language, and English reading and listening skills. Moreover, INVALSI collects information on students’ socioeconomic backgrounds. On the other hand, ANS covers the entire university careers of all the students enrolled in any Italian university since 2008.

The main feature of this integrated dataset is the possibility of quantifying the influence of specific factors on the enrolment choices of Italian graduates. In particular, we are interested in the interplay of three factors: parents’ educational level, the type of high school attended, and the scores in INVALSI maths tests. We consider the cohort of students enrolled in Italian universities in the academic year 2022/23 after graduating from Italian high schools in 2021/22. Students enrolled in healthcare programs are excluded due to the different nature of mobility associated with the enrolment in those programs in Italy. Finally, also students enrolled in online universities are excluded.

5 Modelling

We fit a *mixed-effects logistic model* [13] of the form

$$\text{logit}(\mu_{ij}) = \mathbf{X}_{ij}\beta + \mathbf{Z}_{ij}b_i$$

- $\mathbf{Z}_{ij}$ is the vector of random covariates for the i -th observation in the j -th group. Here, $\mathbf{Z}_{ij}$ are the provinces where students attended high school;
- $b_i \sim \mathbf{N}(0, D)$ is the vector of random effects associated with the $\mathbf{Z}_{ij}$;
- $\mathbf{X}_{ij}$ is the vector of fixed covariates for the i -th observation in the j -th group;
- β is the vector of fixed covariate coefficients.

The response variable identifies *mover* and *stayer* students. Along with the main effect of students' country of origin, we account for the interplay between the high school track, INVALSI maths scores and parents' educational level adding a three-way interaction to the model. Due to the model complexity, the results are reported in terms of median estimated probabilities.

6 Results

This section is devoted to the results of the mixed-effects logistic model. The median estimated probabilities of enrolling in central-northern regions are shown in Fig. 1. On average, students from humanistic licei, with good performance in mathematics, and with tertiary-educated parents are the most likely to leave southern regions. In detail, the interplay of mathematics ability and parents' education seems to be more evident among students from humanistic and scientific

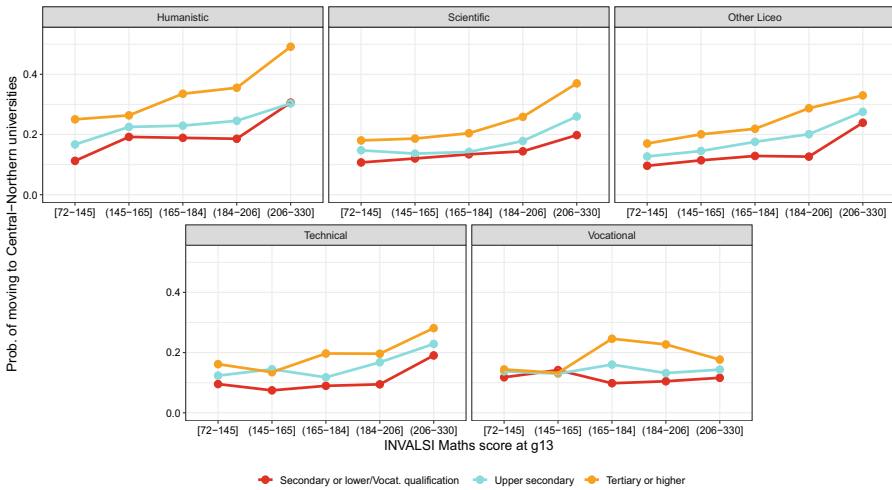


Fig. 1. Median estimated probabilities of moving to Centre-North according to the high school track, parents' educational level, and INVALSI maths score.

schools, with a gap of around 0.2 in the median estimated probability between top-performing students with tertiary-educated parents and their peers with secondary- (or lower-) educated parents. Moreover, it is interesting to observe that the increase in the median estimated probability is stronger for students lying in the last quintile of mathematics scores, highlighting that the best students are the most likely to move to central-northern regions. Finally, caution is advised when interpreting the estimated probabilities for vocational students due to their limited sample size.

Conclusions

This work aims to enrich the literature on inequalities in higher education by providing insights into the interplay of social background, high school track, and academic ability on the mobility choices of students from southern Italy. The results of a mixed-effects logistic model align with our hypotheses. There is a strong positive association between social origin and mobility, as students from more highly educated parents are more likely to migrate from southern regions. Furthermore, the hypothesis of a “boosting effect” is confirmed. Migration to central-northern regions is indeed more likely for students with higher educational achievements; consequently, this exacerbates the loss of human capital in southern regions in favor of central-northern ones, amplifying regional disparities. Finally, the findings underscore the critical necessity of adopting an intersectional approach to deal with educational data. Every stage of education is influenced by a wide range of factors that contribute to shaping the future trajectories of students, starting from their earliest years in school.

Data analysis and manipulation

The data used in this study have been processed in accordance with the research protocol for the study “From high school to the job placement: analysis of university careers and university mobility from Southern to Northern Italy” among the Ministry of University and Research, the Ministry of Education and Merit, the University of Palermo as the lead institution, and the INVALSI Institute. The reference researcher is Massimo Attanasio

References

1. Breen, R., Luijkx, R., Müller, W., Pollak, R.: Nonpersistent inequality in educational attainment: evidence from eight European countries. *Am. J. Sociol.* **114**(5), 1475–1521 (2009)
2. Vergolini, L., Vlach, E.: Family background and educational path of Italian graduates. *High. Educ.* **73**, 245–259 (2017)
3. Mariani, V., & Torrini, R. The geographic divides of the Italian higher education system. Bank of Italy Occasional Paper (675) (2022)

4. Panichella, N.L.: mobilità territoriale dei laureati meridionali: vincoli, strategie e opportunità. *Polis* **23**(2), 221–246 (2009)
5. Impicciatore, R., Tosi, F.: Student mobility in Italy: the increasing role of family background during the expansion of higher education supply. *Res. Soc. Stratification Mobility* **62**, 100409 (2019)
6. Türk, U.: Socio-economic determinants of student mobility and inequality of access to higher education in Italy. *Netw. Spatial Econ.* **19**(1), 125–148 (2019). <https://doi.org/10.1007/s11067-019-09445-w>
7. Attanasio, M., Priulla, A.: Chi rimane e chi se ne va? un'analisi statistica della mobilità universitaria dal mezzogiorno d'Italia. In: Nord, V (ed.) *Le nuove e vecchie rotte delle migrazioni universitarie*, pp. 27–63, Franco Angeli Editore (2020) ISBN: 9788835105626
8. Jackson, M., Erikson, R., Goldthorpe, J.H., Yaish, M.: Primary and secondary effects in class differentials in educational attainment: the transition to A-level courses in England and Wales. *Acta Sociologica* **50**(3), 211–229 (2007)
9. Bernardi, F.: Unequal transitions: selection bias and the compensatory effect of social background in educational careers. *Res. Soc. Stratification Mobility* **30**(2), 159–174 (2012)
10. Contini, D., Salza, G.: Too few university graduates. inclusiveness and effectiveness of the Italian higher education system. *Socio-Econ. Planning Sci.* **71**, 100803 (2020)
11. Panichella, N., Cantalini, S.: Is geographical mobility beneficial? the impact of the south-to-north internal migration on occupational achievement in Italy. *Popul. Res. Policy Rev.* **42**(5), 78 (2023)
12. Database MOBYSU.IT: Mobilità degli studi universitari italiani, Research Protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II, PI: Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA
13. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. arXiv preprint [arXiv:1406.5823](https://arxiv.org/abs/1406.5823) (2014)



A Bayesian Approach to Clustering Using the Proper Bayesian Bootstrap

Federico Maria Quetti^{1(✉)}, Elena Ballante², and Silvia Figini²

¹ Department of Mathematics, University of Pavia, Pavia, Italy
federicomaria.quetti01@universitadipavia.it

² Department of Political and Social Sciences, University of Pavia, Pavia, Italy

Abstract. Few applications of standard bootstrap techniques exist in the literature in clustering problem settings. This paper presents an approach to enhance clustering methods using proper Bayesian bootstrap: in the proposal, prior knowledge about the group densities is introduced, a gaussian mixture model is taken as the prior for the distribution of data.

Our approach is organized in two steps. Efron bootstrap is used to estimate the parameters of the prior. Proper Bayesian bootstrap is used to resample from a mixture of the prior and empirical distribution of data.

Novelty elements and robustness of the proposal are presented as empirical evidence on the Iris dataset.

Keywords: Proper Bayesian Bootstrap · Bayesian Clustering · Bagging · Fuzzy Clustering

1 Introduction

Cluster analysis is the field of machine learning that deals with partitioning data by finding groups in an unsupervised fashion. Clustering techniques are widely used in various applications, where capturing the inherent structure of data plays a pivotal role for the analysis.

Abundant literature exists in the field regarding methods of clustering [1]; yet, being the problem unsupervised, the research on improvements to existing methods is still an open point, leaving room to further developments. In particular, the integration of ML with techniques from statistical learning and Bayesian statistical learning can provide improvements. It has been shown in the literature that the application of bagging techniques to standard clustering methods improves results [2] and brings new information of fuzzy clustering type.

In the following, the aim is to integrate a recent Bayesian bootstrap technique to the problem of clustering, providing a prior knowledge integration scheme. The rest of the paper is structured as follows: Sect. 2 provides the background for the concepts of clustering, bootstrap, and proper Bayesian bootstrap, Sect. 3 is devoted to the description of the methods proposed, in Sect. 4 preliminary results are shown. Conclusions are drawn in Sect. 5.

2 Background

2.1 Clustering

Clustering is the process of grouping data points together based on their similarity with respect to certain features, aiming to uncover inherent patterns or structures within the dataset. The methods differ on the approach chosen and with respect to the similarity definition used.

Several approaches are present in the literature: hierarchical clustering involves organizing data into a tree-like structure, where clusters are nested within one another according to a similarity measure; density-based clustering assigns groups based on regions of high density separated by regions of low density; partition-based clustering, given the number of clusters, creates a partition of data points iteratively, aiming to minimize a cost function.

The results of the following work apply to partition-based clustering methods. Specifically, the focus is on K-means [3], in which the partition is defined so that at every step each point is assigned to the nearest cluster centroid, to minimize the total within sum of squares of the cluster.

2.2 Bootstrap Methods for Clustering

Bootstrap is a statistical resampling technique used to estimate the distribution of a statistic. In the context of ensemble techniques, it provides robust estimates of said statistics: a bagging procedure, where results obtained for each bootstrap replica are aggregated, is used in the proposal, following the results of [2], specifically the algorithm BagClust1 that we briefly outline. Given a learning set L , the number of bootstrap samples B , the chosen number of clusters K , and a clustering method P , the algorithm applies the clustering procedure P to the original learning set L obtaining initial cluster labels. Afterwards, for each bootstrap sample, it applies the clustering procedure again and permutes the cluster labels to maximize overlap with the original clustering. After B iterations, it assigns an aggregated cluster label to each observation based on majority vote, while also recording the participation ratios of votes for each cluster assignment, following a fuzzy clustering logic.

Several kinds of bootstrap methods are present in literature, we mention the most relevant for the work. Efron's bootstrap [4] consists in generating independently each replica from the empirical probability distribution function (pdf) of sample data; equivalently, one can view the procedure as sampling a realization of the cumulative distribution function (cdf), that is drawing a weights vector w for the observations from a Multinomial distribution, and defining

$$F^*(x) = \sum_{i=1}^n \frac{w_i}{n} \mathbb{I}\{X_i < x\} \quad (1)$$

with $(w_1, \dots, w_n) \sim \text{Mult}(n, \frac{1}{n} \mathbf{1}_n)$ and $\mathbb{I}[X_i \leq x]$ indicator function.

An alternative bootstrap procedure, called Bayesian bootstrap [5], can be defined: the proposed method involves sampling Dirichlet weights so to obtain the discrete cdf

$$F^*(x) = \sum_{i=1}^n w_i \mathbb{I}[X_i \leq x] \quad (2)$$

where $(w_1, \dots, w_n) \sim D(\mathbf{1}_n)$.

2.3 Proper Bayesian Bootstrap

In a Bayesian framework, the first approach to prior knowledge integration to the process of resampling data, was proposed in [6]. As the aim of bootstrap is to estimate the true cdf, typically a model for the distribution generating the data is needed. The most widely studied is the Dirichlet process; a Dirichlet process prior has parameter wF_0 , comprised of two quantities: a baseline distribution function F_0 , defining the prior distribution, and a non-negative scaling precision parameter w , which determines how the prior is concentrated around F_0 . The choice of the Dirichlet process is motivated by the following fact [7]: given a random sample $x_1, \dots, x_n$ from $F \sim \mathcal{DP}(wF_0)$, the posterior generating process of the distribution is again a Dirichlet:

$$F|X \sim \mathcal{DP}((w+n)G_n) \quad (3)$$

where

$$G_n = \frac{w}{w+n}F_0 + \frac{n}{w+n}F_n \quad (4)$$

The parameter of the Dirichlet process, given the data, becomes a convex combination of the prior guess F_0 and the empirical cdf F_n .

3 Our Methodological Proposal

The first step of the proposal is based on the application of BagClust1 directly on data. Under the Bayesian paradigm, the second step prescribes the definition of a suitable prior for the generating process underlying bootstrap resampling: in the framework of proper Bayesian bootstrap, a Dirichlet process prior $\mathcal{DP}(wF_0)$ is considered. The confidence w assigned to the prior chosen is a free parameter of the model. The baseline distribution F_0 is imposed as the cumulative of a suitable Gaussian mixture probability density, defined as follows:

$$f_{\theta} = \sum_{j=1}^K p_j f_{\mu_j, \Sigma_j} \quad (5)$$

where $\theta = (p_j, \mu_j, \Sigma_j)$, $j = 1, \dots, K$ are the mixture parameters, and $f_{\mu_j, \Sigma_j} \sim \mathcal{N}(\mu_j, \Sigma_j)$ is the multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j . Knowledge obtained from data drives the estimation of parameters of the mixture model to be imposed as the prior.

The mixture parameters associated to component j are evaluated as follows. The weight p_j , $0 \leq p_j \leq 1$, $\sum_{j=1}^K p_j = 1$, is evaluated as the relative participation to cluster j in the dataset initial cluster labeling. The mean $\boldsymbol{\mu}_j$, representing the cluster centroid, is evaluated as a mean of data points weighted by j th participation ratios recovered from the first step:

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^n PR_j(\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^n PR_j(\mathbf{x}_i)}$$

The variance matrix $\boldsymbol{\Sigma}_j$ is linked to the informativeness of the prior. In the analysis described in Sect. 4 this is defined as $\boldsymbol{\Sigma}_j = s\mathbf{1}_d$, $s > 0$, where d is the dimension of the data points. Upon selection of a value for s in the variance matrix $\boldsymbol{\Sigma}_j = s\mathbf{1}_d$, the prior pdf for the whole dataset is entirely defined.

The second step of the procedure is based on proper Bayesian bootstrap: $m = n$ observations are generated from the convex combination of the defined prior and empirical cdf defined as:

$$G_n = (\omega + n)^{-1}(\omega F_0 + nF_n)$$

where ω is the assigned confidence parameter. Then, we assign weights $\omega_1^b, \dots, \omega_m^b$, drawn from $D(\frac{\omega+n}{m}, \dots, \frac{\omega+n}{m})$, to these observations. The proper Bayesian bootstrap resample presents newly sampled values as well as original dataset values, which are the focus of the cluster labels assignment. Iteratively, K-means is applied to the B resampled learning sets. Participation ratios for the original data points are re-evaluated.

The procedure, from the fuzzy assignment to the clusters, gives an aggregated value of the cluster label for the original data points as: $\operatorname{argmax}_{1 \leq k \leq K} PR_k(\mathbf{x}_i)$.

4 Preliminary Results

In order to present the advantages of the proposed method, results obtained on the Iris dataset are shown: the results are presented as contingency tables where rows represent true clustering labels while columns the predicted ones.

In this work, the number of clusters K is considered a known parameter; methods for adaptively choosing K will be analyzed in further research.

The extensively studied Iris dataset is comprised of 150 data points under 4 features; it comes with the true cluster labels of the points, to be tested against the assignment of the clustering procedure.

Against the benchmark results for K-means, shown in Table 1, our proposal is tested with different choices of parameter; results for different values of the variance parameter s , and of the confidence w , are shown in Table 2a and Table 2b, respectively.

From Table 2b, we see that the method is robust when the prior has variance $s = 1$, meaning that the model imposed for the dataset pdf is well posed. The results on the sensitivity analysis on the variance described in Table 2a show that increasing s the cluster structure is not captured well, worsening the performance in cluster 2. These results confirm the necessity of an accurate tuning of model parameters.

Table 1. Results for K-means clustering without bootstrap

	1	2	3
1	50	0	0
2	0	48	2
3	0	14	36

Table 2. Results for the proposed method

	1	2	3	<i>s</i>	<i>w</i>
1	50	0	0	10.1	
2	0	47	3	10.1	
3	0	13	37	10.1	
1	50	0	0	100.1	
2	0	46	3	100.1	
3	0	15	35	100.1	
1	50	0	0	1000.1	
2	0	1	247	1000.1	
3	0	0	50	1000.1	

(a) Contingency tables for different choices of variance parameter *s*

	1	2	3	<i>s</i>	<i>w</i>
1	50	0	0	10.3	
2	0	46	4	10.3	
3	0	13	37	10.3	
1	50	0	0	10.5	
2	0	45	5	10.5	
3	0	10	40	10.5	
1	50	0	0	10.7	
2	0	45	5	10.7	
3	0	9	41	10.7	

(b) Contingency tables for different choices of confidence *w*

5 Conclusions

In this work we proposed a new clustering approach, based on the deployment of bagging techniques. With little computational effort, the method adds the benefit of a Bayesian interpretation of the data generating process, while at the same time having the potential to retrieve additional information of participation while preserving (and in some cases bettering) the benchmark of usual methods. Future directions of work include the usage of different methods from K-means, more refined choices of the prior, and the extension of the method to other types of aggregation used in a cluster setting (BagClust2, [2]). Further analysis will be also carried out for diverse dataset, focusing particularly on dimensionality effects.

References

1. Anil, K.J., Narasimha Murty, M., Flynn, P.J.: Data clustering: a review. *ACM Comput. Surv. (CSUR)* **31**(3), 264–323 (1999)

2. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* **19**(9), 1090–1099 (2003)

3. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**, 768–769 (1965)

4. Efron, B.: Bootstrap methods: another look at the jackknife. In: *Breakthroughs in Statistics: Methodology and Distribution*, pp. 569–593. Springer (1992). https://doi.org/10.1007/978-1-4612-4380-9_41

5. Rubin, D.B.: The bayesian bootstrap. *The annals of statistics*, pp. 130–134. JSTOR (1981)
6. Muliere, P., Secchi, P.: Bayesian nonparametric predictive inference and bootstrap techniques. *Annals Instit. Stat. Math.* **48**, 663–673 (1996)
7. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230. JSTOR (1973)



Analysis of Higher Education Dropouts Dynamics Through Functional Decomposition of Recurrent Events on Time-to-Event Processes

Alessandra Ragni^(✉), Chiara Masci, and Anna Maria Paganoni

MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci
32, 20133 Milan, Italy

{alessandra.ragni, chiara.masci, anna.paganoni}@polimi.it

Abstract. In this paper, we delve into the dynamics of higher education dropouts through a novel approach that combines time-to-event processes analysis with multilevel functional decomposition of recurrent events. Utilizing data from Politecnico di Milano's administrative archive, we explore dropout patterns across bachelor's degree programs and schools. By leveraging models for recurrent events and employing functional principal component analysis techniques, our methodology expands upon existing frameworks to accommodate hierarchical data structures. Through simulations and analysis, we demonstrate the efficacy of our approach in reconstructing dropout dynamics, shedding light on critical periods, disentangling the effects given by degree programs and schools and informing proactive strategies for educational institutions.

Keywords: Students Dropout · Compensators Decomposition · Multilevel Principal Component Analysis · Recurrent Events

1 Introduction

The higher education system worldwide faces a common challenge: high dropout rates, especially among bachelor's students [1]. This problem is particularly pronounced in Italy, where even more significant percentages of students leave their studies, often within the first two years of enrollment [2]. These dropout rates represent losses not only in terms of money and time but also in educational opportunities. Understanding the patterns of this trend is crucial for addressing it effectively.

In this work, we focus on data sourced from the administrative archive of Politecnico di Milano. We aim to explore the time-dependent pattern of dropout rates across the three years of bachelor's programs, examining the number of dropouts for each degree program nested within each school. Our methodology

is inspired by [3,4], where models for recurrent events and functional data analysis techniques are leveraged, but we extend and generalize the framework to accommodate hierarchical data structures.

While existing studies, both within Politecnico di Milano’s administrative data [5] and globally, have attempted to address the impact of grouping factors (e.g., degree programs) on time to dropout within three years after enrollment using shared frailty Cox proportional hazard models, our method presents a novel approach. We offer a more nuanced analysis of dropout behaviour over time, specifically honing in on the degree program and school dynamics.

2 Dataset

The data employed in this study are sourced from the administrative archive¹ of Politecnico di Milano, encompassing academic careers of students enrolled in Bachelor’s degree programs.

For our analysis, we focused specifically on bachelor’s students enrolled in the academic year 2017 who did not change their degree program. We excluded students of online bachelor’s programs and focused on dropout occurrences that took place on distinct days between October 1, 2017, and March 10, 2021, (i.e., 1256 days, which correspond to 7 semesters), basing this choice on previous studies’ insights. Following preprocessing, our dataset consists of 1008 distinct dropout days, encompassing 25 distinct degree programs (denoted in the following Section as *units*) across 4 different schools² (denoted in the following Section as *clusters*).

3 Methods

3.1 Model Formulation for Recurrent Events

Observations may be considered as the realizations of n processes, where N_{ij} is the stochastic process which counts the observed events related to the j^{th} unit for $j = 1, \dots, n_i$ and i^{th} cluster for $i = 1, \dots, I$ and $\sum_{i=1}^I n_i = n$. Doob-Meyer decomposition theorem [6] states that $M_{ij}(t) = N_{ij}(t) - \Lambda_{ij}(t)$ is a zero-mean, uniformly integrable martingale, where $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(s) ds$ is the cumulative hazard function, also called *compensator* in this setting, and $\lambda_{ij}(t)$ the conditional intensity function.

Following Sect. 3.1 in [4] for marked point processes, the conditional intensity function can be seen as dependent on a mark³ ω_{ij} , i.e., $\lambda_{ij}(t, \omega_{ij}) = \lambda_{ij,g}(t) \cdot$

¹ Universities systematically collect a variety of data, including comprehensive career information, such as start and end dates, as well as dropout occurrences (if they occurred) and details regarding the degree program in which the student is enrolled, along with any eventual changes made during their academic journey.

² Indicated as **sA**, **sB**, **sC** and **sC**.

³ A vector, whose components are magnitudes of information associated to each jump time.

$f_{ij}(\boldsymbol{\omega}_{ij})$ where $\lambda_{ij,g}(t)$ is the intensity process of the counting process (ground intensity) and f_{ij} is the multivariate density of the marks $\boldsymbol{\omega}_{ij}$.

Making use of the so-called Andersen Gill (AG) model [7] for recurrent events, under the assumption that f_{ij} depends on $\mathbf{z}_{ij}(t)$ (that are some time-dependent features related to the marks $\boldsymbol{\omega}_{ij}$ for unit j in cluster i), we get

$$\begin{aligned}\lambda_{ij}(t) &= \lambda_{ij}(t, \boldsymbol{\omega}_{ij}) = Y_{ij}(t) \lambda_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_{ij}(t)\} \exp\{\boldsymbol{\theta}^T \mathbf{z}_{ij}(t)\} \\ &= Y_{ij}(t) \lambda_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_{ij}(t) + \boldsymbol{\theta}^T \mathbf{z}_{ij}(t)\}\end{aligned}\quad (1)$$

where $\mathbf{x}_{ij}(t)$ are the (time-dependent) vectors of covariates of the j^{th} unit in i^{th} cluster, $\lambda_0(t)$ is the baseline hazard function shared across clusters and units and Y_{ij} takes the roles of the censoring variable (i.e. assumes value 1 when unit j in cluster i is under observation).

3.2 Compensators Reconstruction and Decomposition

Reconstruction. $\hat{\Lambda}_0(t)$ is estimated as a step-function using the Breslow estimator [8] and then smoothed as described in [3] obtaining $\hat{\Lambda}_0(t)$ and treated as a functional data object [9]. Being $[t_k^{(ij)}, t_{k+1}^{(ij)}]$ the intervals whose extremes are the jump times for each unit j in cluster i , $\hat{\Lambda}_{ij}(t)$ can be estimated as follows:

$$\begin{aligned}\hat{\Lambda}_{ij}(t) &= \int_0^t Y_{ij}(s) \lambda_0(s) \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}(s) + \hat{\boldsymbol{\theta}}^T \mathbf{z}_{ij}(s)) ds \\ &= \sum_{k=0}^{N_{ij}(t)} \int_{t_k^{(ij)}}^{\min(t_{k+1}^{(ij)}, t)} \lambda_0(s) \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}(s) + \hat{\boldsymbol{\theta}}^T \mathbf{z}_{ij}(s)) ds \\ &\simeq \sum_{k=0}^{N_{ij}(t)} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}(t_k^{(ij)}) + \hat{\boldsymbol{\theta}}^T \mathbf{z}_{ij}(t_k^{(ij)})) \int_{t_k^{(ij)}}^{\min(t_{k+1}^{(ij)}, t)} \lambda_0(s) ds \\ &= \sum_{k=0}^{N_{ij}(t)} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}(t_k^{(ij)}) + \hat{\boldsymbol{\theta}}^T \mathbf{z}_{ij}(t_k^{(ij)})) [\tilde{\Lambda}_0(\min(t_{k+1}^{(ij)}, t)) - \tilde{\Lambda}_0(t_k^{(ij)})].\end{aligned}\quad (2)$$

Decomposition. $\hat{\Lambda}_{ij}(t)$ are then decomposed making use of Multilevel Functional Principal Component Analysis (MFPCA) [10, 11]:

$$\hat{\Lambda}_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t) + \epsilon_{ij}(t) \quad (3)$$

$$\simeq \mu(t) + \sum_{k=1}^{K_1} \xi_{ik} \phi_k^{(1)}(t) + \sum_{l=1}^{K_2} \zeta_{ijl} \phi_l^{(2)}(t) + \epsilon_{ij}(t) \quad (4)$$

where $\mu(t)$, $\phi_k^{(1)}(t)$, $\phi_l^{(2)}(t)$ are fixed functional effects, with $\phi_k^{(1)}(t)$ and $\phi_l^{(2)}(t)$ being respectively level 1 (i.e., cluster level) and level 2 (i.e., unit level) eigenfunctions, and being ξ_{ik} and ζ_{ijl} level 1 and 2 principal component scores. $\mu(t)$

is assumed as a fixed function while $Z_i(t)$ and $W_{ij}(t)$ are mean 0 stochastic processes (uncorrelated between each other); ϵ_{ij} is observed only when functional data are observed with error.

4 Simulation Study

Our objective is to demonstrate the effectiveness of our proposed compensator reconstruction method, employing AG models, in accurately restoring the simulated shapes derived from time-series data. Data are generated by Non-Homogeneous Poisson Processes extracted by unit-level intensities exhibiting specific similarities within clusters. To accomplish this, we simulate intensities $\lambda_{ij}(t)$ following a similar methodology as described in [10,11], where a one-way functional ANOVA model as Eq. (3) is employed. Subsequently, we integrate $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du$, to obtain a compensator-like shape. Concurrently, NHPPs are generated from the $\lambda_{ij}(t)$, AG models fitted and the compensators $\hat{\Lambda}_{ij}(t)$ reconstructed.

Our simulation study [12] demonstrates the efficacy of the presented method in reconstructing clusters and units time patterns, despite some information loss inherent in the NHPP generation.

5 Results

The selected features for the AG model for recurrent events describing the dropouts are `enum` (i.e. the number of past events), `N_stud` (i.e., the total number of students that enrolled into that degree program in 2017) and `dropout.count` (i.e., the mark of the point process, indicating the number of dropouts occurred in that specific day), which all resulted to be significant (hazard ratios are respectively 1.008, 1.001 and 0.931).

Compensators are then reconstructed as described in Eq. (2). In Fig. 1, we report (on two different scales) the baseline cumulative hazard and the reconstructed compensators. It is interesting to notice the behaviour of the curves: in particular, we can see a steep increase of the dropouts counts at the end of the first year, very accentuated for specific degree programs.

Furthermore, the compensators are decomposed as in Eq. (4) and results are reported in Fig. 2. The first Principal Components (PCs) (panels (a) and (b)) distinguish schools and degree programs with different dropout risks, while the second ones (panels (c) and (d)) distinguish schools and degree programs with different distribution in time of the dropouts. It is interesting to notice that, as an element of novelty, we are able to disentangle the effects given by the schools with respect to the degree programs. In fact, the two PCs relative to schools (level 1) differentiate both the intensities and the dynamics of dropouts over time.

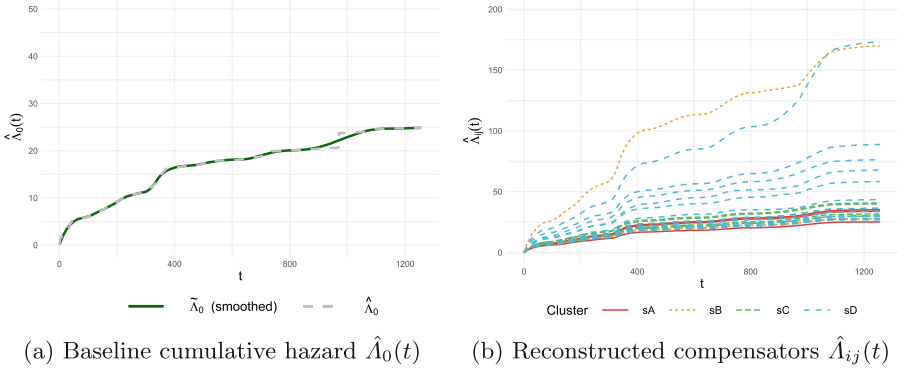


Fig. 1. In panel (a) we show the baseline cumulative hazards of the AG model for recurrent events for the marked stochastic processes describing the dropouts. In panel (b) we represent the reconstructed compensators of the latter processes, each line representing a different degree program (unit j) and each color representing a different school (cluster i), as indicated in legend.

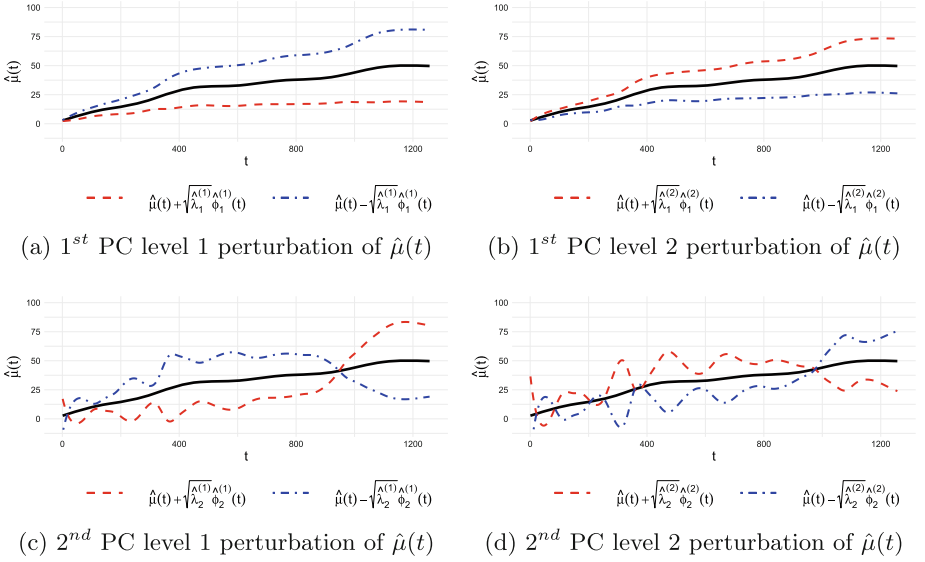


Fig. 2. Average compensators curves $\hat{\mu}(t) = \frac{1}{n} \sum_{i,j} \hat{\Lambda}_{ij}(t)$ and their perturbations as indicated in legends, for 1st and 2nd Principal Components (PCs) for level 1 in left panels (a) & (c) and for level 2 in right panels (b) & (d).

6 Conclusion and Discussion

In this paper, we focus on dropout occurrences along time within each degree program and we introduce a novel method for (i) constructing the compensators

of recurrent events and (ii) through MFPCA, decomposing them to discern the principal driving components and behaviour in time at both the degree program and school levels.

Our findings yield valuable insights for universities seeking to understand peak dropout periods throughout the students' bachelor, thereby enabling the development of targeted policies aimed at reduction or anticipation of dropout rates, informing proactive strategies for educational institutions.

Further developments of this work will include additional simulations, integration of multi-year enrollment data, comparison of dropout trends before and after the COVID-19 pandemic and integration of the history (contribution of degree program and school) up to a certain time for the prediction of the dropout at student-level.

References

1. OECD Indicators. Education at a Glance 2016 OECD Indicators (2013). https://uned.kneu.edu.ua/wp-content/uploads/2_news.pdf
2. Anvur. Rapporto biennale sullo stato del sistema universitario e della ricerca (2018). <https://www.anvur.it/attivita/rapporto-biennale/rapporto-biennale-2018/>
3. Baraldo, S., Ieva, F., Paganoni, A.M., Vitelli, V.: Outcome prediction for heart failure telemonitoring via generalized linear models with functional covariates. *Scand. J. Stat.* **40**(3), 403–416 (2013)
4. Spreafico, M., Ieva, F.: Functional modeling of recurrent events on time-to-event processes. *Biom. J.* **63**(5), 948–967 (2021)
5. Masci, C., Cannistrà, M., Mussida, P.: Modelling time-to-dropout via shared frailty Cox models. A trade-off between accurate and early predictions. *Stud. High. Educ.* 1–19 (2023)
6. Meyer, P.A.: A decomposition theorem for supermartingales. *Illinois J. Math.* **6**, 193–205 (1962)
7. Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982)
8. Breslow, N.E.: Analysis of survival data under the proportional hazards model. *Int. Statist. Rev./Revue Internationale de Statistique* **43**(1), 45–57 (1975)
9. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. Springer Series in Statistics (2005). <https://doi.org/10.1007/b98888>
10. Di, C.-Z., Crainiceanu, C.M., Caffo, B.S., Punjabi, N.M.: Multilevel functional principal component analysis. *Annal. Appl. Statist.* **3**(1), 458 (2009)
11. Cui, E., Li, R., Crainiceanu, C.M., Xiao, L.: Fast multilevel functional principal component analysis. *J. Comput. Graph. Stat.* **32**(2), 366–377 (2023)
12. Ragni, A., Masci, C., Paganoni, A.M.: Analysis of Higher Education Dropouts Dynamics through Multilevel Functional Decomposition of Recurrent Events on Time-to-Event Processes. In progress



The PublicWorksFinanceIT R Package: Retrieve and Visualize Public Works Data

Lorena Ricciotti^(✉) 

Università degli Studi di Bari Aldo Moro, Bari, Italy
lorena.ricciotti@uniba.it

Abstract. The ‘PublicWorksFinanceIT’ package facilitates the retrieval and analysis of financial data related to public works in Italy, with a specific focus on soil defense investments. The data is gathered from three main sources: the OpenCoesione website, OpenBDAP website, and the ReNDiS database. This package provides a user-friendly tool for accessing and analyzing the data, ensuring real-time updates. The records from these sources are linked using unique project codes, preventing duplication. The data is geographically referenced, allowing users to obtain information at the region, province, and municipality levels. The package includes functions for data retrieval, and visualization on maps.

Keywords: Soil defense investments · public works · R package · data retrieving · geospatial data

1 Introduction

In recent years, the proliferation and diversity of publicly accessible data sources on the web has experienced exponential growth. This surge is fueled by both the digitalization of information and an increasing expectation for entities, both public and private, to share their data widely. Recognizing this trend, the European Union implemented the Digital Decade Policy Programme (DDPP) on January 9, 2023, aiming to advance digital services, strengthen cyber-security, and promote technologies such as artificial intelligence, cloud computing, and big data in the business sector.

Despite the availability of open data across various research fields, challenges persist, including inconsistent organization and format disparities among datasets [1, 2]. To address these issues, the FAIR approach has been introduced, focusing on Findability, Accessibility, Interoperability, and Reusability of data [3]. A wide range of efforts have been done to provide clean data [4–6]. Examples are the **webchem** package [7] for retrieving chemical datasets, the **ARPALData** package [8] for retrieving and visualizing environmental data.

Building on these precedents, the ‘PublicWorksFinanceIT’ package was developed by the author with the aim of facilitating the access and visualization of financial data related to public works in Italy [9]. The dataset is meticulously curated from diverse sources, aligning with FAIR data principles. Specifically

tailored to soil defense investments, the package contributes to environmental protection efforts by ensuring the longevity and sustainability of public works. Financial data in this context have proven crucial for assessing the effectiveness of policies and actions, as demonstrated in studies evaluating public biodiversity investments and storm surge impacts [10,11].

Data are retrieved from three different platforms, namely the OpenCoesione, OpenBDAP, and ReNDiS. These sources offer comprehensive insights into investments made for public works, providing details on project codes, nature of interventions, and financial allocations. Furthermore, the dataset is enriched with geographical references, enhancing its utility for researchers and policymakers alike.

2 Public Works and Soil Defense

Public works refer to infrastructure projects and services financed and carried out by government authorities to meet the collective needs of society. This includes a variety of facilities like roads, bridges, public transit, water treatment plants, energy infrastructure, public buildings, and other essential structures. These projects aim to improve community well-being by fostering economic development, ensuring public safety, and enhancing residents' quality of life. Specifically, public works also include soil defense interventions, which play a crucial role in promoting sustainable development. These initiatives encompass a variety of interventions designed to reduce the risks associated with environmental factors like erosion, landslides, and flooding. An example of public works, which specifically includes soil defense interventions crucial for sustainable development, is highlighted in the context of addressing coastal erosion risks. Coastal regions, indeed, face increased threats from climate change-induced hazards. Mitigating these risks requires substantial investments in soil defense measures.

3 Methodology

The package utilizes data from three key repositories: OpenCoesione, OpenBDAP, and ReNDiS. The OpenCoesione and OpenBDAP datasets provide insights into public works investments, while ReNDiS focuses specifically on soil defense financing. Data can be sourced and visualized through the use of 11 different functions which can be classified as: functions for data/information retrieval, and functions for data visualization.

3.1 Data Sources and Geospatial Referencing

The OpenCoesione, OpenBDAP, and ReNDiS datasets offer valuable information on public investments, timelines, and soil defense initiatives.

The 'Open Coesione' website (<https://opencoesione.gov.it/it/>) utilizes data from Cohesion Policy, a commitment by EU member states to deliver more cohesive programs and data at national and European levels. This resource undergoes

bi-monthly updates, providing a comprehensive perspective on public investments sourced from Italian and European funding channels. Data is gathered at the municipal level, showcasing the resilience measures implemented by each municipality alongside the corresponding budget allocations. This database offers a detailed insight into economic investments related to public infrastructures, compiling data from national and European programs. It presents a thorough breakdown of financing allocation between national and European resources, specifying the total funding and individual contributions from each source.

The ‘OpenBDAP’ database (<https://openbdap.rgs.mef.gov.it/>) is administered by the Italian Ministry of Economy and Finance, specifically by the *Ragioneria Generale dello Stato*. While it primarily houses data related to the Italian Public Administration, for the purpose of the package it focuses on the dataset related to investments made for public works, with a particular emphasis on soil defense. This dataset provides comprehensive details on the utilization of public funds in Italy across various sectors. It includes information about interventions such as their nature, CUP (Unique Project Code), investment sectors, intervention descriptions, implementation timelines, sub-sector details, and other relevant attributes. From a financial perspective, the dataset outlines projected costs of interventions and allocates financial flows among national, local authorities, and private entities.

The Repertorio Nazionale degli Interventi per la Difesa del Suolo (ReNDiS) database (<http://www.rendis.isprambiente.it/rendisweb/>), provided by ISPRA, serves as a repository for data related to soil defense investments. It offers users valuable insights, providing details on the total resources allocated to each project, the specific category of intervention, and the corresponding CUP. While leveraging the comprehensive MEF database as its primary source, ReNDiS stands out through its ongoing implementation, emphasizing a commitment to continuous improvement and the seamless integration of the latest data. This is facilitated by direct access granted to institutions, enabling them to upload project data directly to the platform.

The three databases have been enriched with crucial geographical references at the municipal level. Two types of georeferencing are available: either the coordinates of the centroids of each municipality or the areal data, depicting the polygons outlining the boundaries of each municipality. These options are extracted from the ISPRA database through dynamic queries. ISPRA takes this data from the ISTAT database ensuring not only flexibility but also a guarantee of continuous updates.

In the ReNDiS database, georeferencing seamlessly integrates during the data download via queries, capitalizing on the existing georeferences within the database. For the Open Coesione dataset, centroids or polygons are linked through province and municipal codes. In contrast, the Open BDAP datasets lack inherent geographical references. However, a dedicated localization dataset associates CUPs with specific region, province, and municipal codes. Following this association, georeferencing is then appended using province and municipal codes.

3.2 Functions in the PublicWorksFinanceIT Package

Users can access clean data from three distinct sources through the use of specialized functions. Each source corresponds to a specific function for data retrieval. Specifically, the `get_data_RENDIS()` function is designed for the ReNDiS database, the `get_data_BDAP()` function for OBDAP, and there are two distinct functions for Open Coesione. The `get_data_region_OC()` function retrieves data based on the user's specified region of interest, while the `get_data_theme_OC()` function sources data according to specific themes related to public works.

Additionally, three auxiliary functions are available to assist users in obtaining information about the parameters to use in the main functions, tailored to their interests. The `get_codes()` function provides users with ISTAT codes for regions, provinces, and municipalities, simplifying the process of finding corresponding codes for the specific areas of interest. Moreover, the `get_type_RENDIS()` function furnishes a list of intervention types in soil defense as presented in the ReNDiS database. This allows users to filter data retrieval based on their interest in specific types of interventions. Last, the `get_info_OC()` function is designed for OpenCoesione. Since it doesn't use ISTAT codes for Italian regions, this function aids in retrieving the different codes associated with each region. Moreover, users can obtain information on the codes linked to each project's theme, facilitating their use in the function `get_data_theme_OC()` for data retrieval based on themes.

Concerning visualization functions, three distinct types of plots are available. The first is a bar plot, useful for illustrating the various fund allocations in a specific region. This can be achieved using the `plot_funds_bar()` function, where users specify the financial variables of interest.

The package also provides two map visualization options, leveraging georeferencing in the data. If users prefer areal data representation, the `plot_funds_map()` function can be employed. In this case, users can specify a variable of interest and observe its distribution across municipalities, visualizing a map color-coded based on the financial imports. Lastly, for those requesting centroids of municipalities, the `plot_funds_points()` function can be utilized. This function displays centroids for each municipality as circles, with variations in colors and dimensions corresponding to the respective financial amounts.

In conclusion, the `merge_data()` function facilitates users in merging the datasets obtained from the three distinct sources. This results in a comprehensive dataset encompassing 32 key variables, providing users with a complete overview of soil defense investments across their region of interest.

3.3 Example

Here a specific example of soil defense data is presented to show the functionality of the package. Particularly it refers to the Lazio region. To get data from the three different platforms regarding only soil defense interventions the three

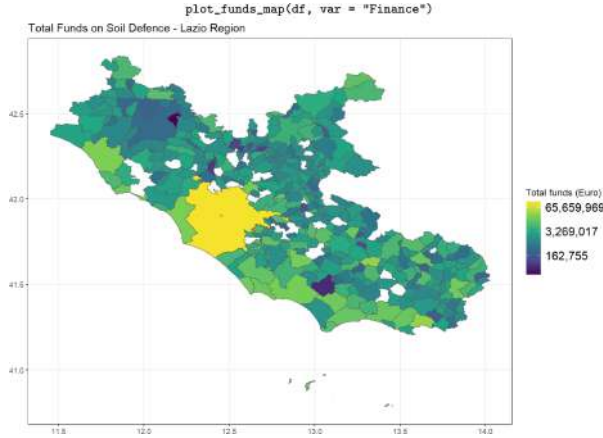


Fig. 1. Map displaying each municipality in the Lazio region, color-coded to represent the corresponding amount of total finance received

functions aforementioned to retrieve data have been used specifically requesting areal data.

Looking at Fig. 1 the municipality that has received the majority of funds is Rome, as indicated by the color of the area. White polygons represent municipalities which do not appear in the dataset. Moreover, as the dataset focuses on soil defense interventions, our analysis highlights a trend: primary funds are predominantly allocated to municipalities along or in proximity to the coastline, whereas inland municipalities seem to receive comparatively fewer allocations.

4 Conclusion

This paper presents a comprehensive analysis of the interplay between public works and soil defense initiatives in Italy. The incorporation of geospatial data enriches the study, providing a vital layer of information for a nuanced comprehension of geographical patterns and distribution of investments. These findings serve as a valuable resource for informed decision-making, underscoring the significance of strategic investments in soil defense within the broader framework of sustainable development. The amalgamation of data from diverse sources, such as the ReNDiS, Open Coesione, and Open BDAP databases, contributes to a holistic understanding of soil defense endeavors, fostering a more resilient and environmentally conscious approach to public infrastructure development. This research not only advances our understanding of the current landscape but also lays the foundation for future endeavors in optimizing resource allocation and promoting sustainable practices in soil defense initiatives. Looking ahead, the insights derived from this package lay the groundwork for future applications and advancements in the field of soil defense initiatives. The comprehensive dataset amalgamated from the ReNDiS, Open Coesione, and Open BDAP databases

provides a robust foundation for continued research and analysis. Future studies could explore more granular aspects of soil defense projects, delving into specific intervention types, regional disparities, and the long-term impact of investments. Moreover, the findings presented in this paper are not only valuable for policymakers and researchers but also hold significance for various stakeholders involved in environmental sustainability and infrastructure development. Local authorities, government agencies, and environmental organizations can leverage this information for strategic planning, identifying areas that require targeted interventions and optimizing resource allocation for soil defense projects. In essence, the ‘PublicWorksFinanceIT’ opens avenues for diverse users, including policymakers, researchers, environmentalists, and local authorities, to collaboratively work towards the shared goal of fostering sustainable development and resilience in soil defense initiatives. The continuous updates and dynamic nature of the databases ensure that future applications can build upon this foundation, contributing to an evolving and adaptive framework for addressing the challenges posed by soil erosion and related environmental concerns.

References

1. Reichman, O.J., Jones, M.B., Schildhauer, M.P.: Challenges and opportunities of open data in ecology. *Science* **331**(6018), 703–705 (2011)
2. Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R.S.: Socio-technical impediments of open data. *Electron. J. e-Gov.* **10**(2), 156–172 (2012)
3. Scheffler, M., et al.: Fair data enabling new horizons for materials research. *Nature* **604**(7907), 635–642 (2022)
4. Farese, M., Soncin, S., Robb, J., Fernandes, R., Tafuri, M.A.: The mediterranean archive of isotopic data, a dataset to explore lifeways from the neolithic to the iron age. *Sci. Data* **10**(1), 917 (2023)
5. Rühr, P.T., Edel, C., Frenzel, M., Blanke, A.: A bite force database of 654 insect species. *Sci. Data* **11**(1), 58 (2024)
6. Thorn, J.P., Bignoli, D.J., Mwangi, B., Marchant, R.A.: The African development corridors database: a new tool to assess the impacts of infrastructure investments. *Sci. data* **9**(1), 679 (2022)
7. Szöcs, E., Stirling, T., Scott, E.R., Scharmüller, A., Schäfer, R.B.: Webchem: an R package to retrieve chemical information from the web. *J. Stat. Softw.* **93**, 1–17 (2020)
8. Maranzano, P., Algieri, A.: Retrieving and Analyzing Air Quality and Weather Data from ARPA Lombardia (2023). <https://CRAN.R-project.org/package=ARPALData>, R package version 1.5.0
9. Ricciotti, L., Pollice, A.: PublicWorksFinanceIT: Soil Defense Investments in Italy: Data Retrieval, Analysis, Visualization (2024). <https://CRAN.R-project.org/package=PublicWorksFinanceIT>, R package version 0.2.0
10. Seidl, A., Mulungu, K., Arlaud, M., van den Heuvel, O., Riva, M.: The effectiveness of national biodiversity investments to protect the wealth of nature. *Nat. Ecol. Evol.* **5**(4), 530–539 (2021)
11. Prah, B.F., Boettle, M., Costa, L., Kropp, J.P., Rybski, D.: Damage and protection cost curves for coastal floods within the 600 largest European cities. *Sci. data* **5**(1), 1–18 (2018)



Oracle Priors in Bayesian Inference: a Frequentist View

Stefano Rizzelli^(✉)

University of Padova, Padova, Italy
stefano.rizzelli@unipd.it

Abstract. In numerous Bayesian procedures, especially for high-dimensional models, certain prior choices induce posterior distributions with optimal properties, according to some frequentist criteria. This note offers a unified treatment of such priors through the notion of oracle priors. Several examples from the most recent literature are considered, illustrating the impact of these priors on statistical inference in specific cases.

Keywords: posterior contraction rates · empirical Bayes · hierarchical Bayes · multiple testing

1 Introduction

Inference within a Bayesian approach with statistical model $\{P_\theta, \theta \in \Theta\}$ for observable Y moves from the specification of a prior distribution on model's parameter θ . An objective approach admits the existence of a true (unknown) distribution and regards the prior as a probabilistic means of expressing one's uncertainty with respect to reality. In this approach, it is natural to question the robustness of posterior inference with respect to different choices of the prior and its (asymptotic) accuracy for increasing levels of information. The perspective adopted is thus inherently frequentist: while one wants to benefit from the advantages of the Bayesian statistical learning machinery (e.g., simultaneous estimation of several data distribution's functionals, use of predictive distributions, etc.), he/she does not waive the guarantees on the quality of inference typically formulated in terms of the actual distribution of the observables.

Several priors have been expressly devised to produce posterior or predictive distributions with some desired frequentist properties. For example, probability matching priors are designed such that the corresponding posterior probabilities of certain regions (approximately) coincide with their coverage probabilities. Similarly, predictive probability matching priors are constructed to match posterior predictive probabilities with the corresponding frequentist probabilities. This note offers a survey of prior distributions whose definition is based on knowledge of the true model's parameter, named therefore *oracle priors*, and which yield posterior distributions with some optimality properties, according to frequentist criteria.

1.1 Scope, Outline and Notation

The main point raised in the sequel is that the notion of oracle prior, in its various facets, provides a key to analysing and understanding frequentist performance achievable with a Bayesian procedure in a wide variety of contexts. Albeit desirable, oracle priors are not usable in practice. Our aim is thus to take stock of situations in which it is still possible to construct a prior leading to similar posterior inferential performance, offering an overview of recent advances in frequentist-Bayes literature. In most of the considered cases, observables are of sequential form $Y = (Y_1, \dots, Y_n)$ and a prior is chosen within a given class $\{\Pi_\lambda, \lambda \in \Lambda\}$ to carry out inference through the posterior distribution

$$\Pi_\lambda^{(n)} = \Pi_\lambda(\cdot \| Y);$$

more generally, n stands for an index whose increase denotes accruing data information. Oracle priors thus correspond to an optimal choice λ^* of the hyperparameter to which the prior distribution is indexed. Typically, an inferential performance similar to those of posterior distributions originated by an oracle prior is obtainable through *empirical Bayes* posteriors, derived by plugging a given dependent choice $\hat{\lambda}_n$, i.e.

$$\Pi_{\hat{\lambda}_n}^{(n)}(B) = \Pi_\lambda^{(n)}(B)|_{\lambda=\hat{\lambda}_n},$$

or with appropriate hierarchical posteriors, derived by assigning a hyper prior $\lambda \sim H$. In the particular case of unidimensional parameter models, the effects of an oracle prior selection on the higher order properties of estimation are illustrated, as well as the interplay between oracle and empirical Bayes.

2 Estimation

2.1 Finite-Dimensional Models

We first consider dominated finite-dimensional, parametric models $\{p_\theta^{(n)}, \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ and $d \in \mathbb{N}_+$. In this setting, for regular models and priors, posterior distributions share same first-order asymptotic properties, e.g. contraction rates are typically of the form $M_n/\sqrt{n}$, for any $M_n \rightarrow \infty$ such that $M_n = o(\sqrt{n})$, no matter what the specific form of the prior is. To unravel optimality properties of some prior distribution in the induced posterior estimation procedures, it is thus necessary to consider finer, higher-order asymptotic properties. Below some intuition is provided concerning higher order accuracy of point estimation using posterior means.

We assume herein that the chosen prior has density $\pi_\lambda(\theta)$. [4] define the corresponding oracle set of hyperparameters as

$$\Lambda^* = \operatorname{argmax}_{\lambda \in \Lambda} \pi_\lambda(\theta_0).$$

Intuitively, Λ^* is the set of hyperparameter choices λ^* which mostly favour the true parameter value *a priori*. A new, information-theoretic characterisation has been recently provided by [5], who show that (loosely speaking) oracle values λ^* are those

which asymptotically minimise the Kullback-Leibler divergence from marginal likelihoods to the true density $\text{KL}(p_{\theta_0}^{(n)} \| p_{\Pi_\lambda}^{(n)})$, where $p_{\Pi_\lambda}^{(n)}(\cdot) = \int_{\Theta} p_\theta^{(n)}(\cdot) d\Pi_\lambda(\theta)$. See Theorem 2.1, Corollary 2.2. and Corollary 2.3 in [5] for details. It is also known that empirical Bayes selection of hyperparameters through maximization of the marginal likelihood (MMLE)

$$\hat{\lambda}_n = \operatorname{argmax}_{\lambda \in \Lambda} p_{\Pi_\lambda}^{(n)}(Y_1, \dots, Y_n)$$

asymptotically attains the oracle set Λ^* ; see [4], Theorem 2, for first rigorous results and [5], Theorem 2.3, for recent generalisations to much wider setups.

We now focus on Laplace regular models [2] with MLE satisfying $\hat{\theta}_n = \theta_0 + O_{\mathbb{P}}(1/\sqrt{n})$ and, assuming for simplicity that $d = 1$,

$$-\frac{\partial^2}{\partial \theta^2} \frac{\log p_\theta^{(n)}(Y_1, \dots, Y_n)}{n} \Big|_{\theta=\theta_0} = I(\theta_0) + o_{\mathbb{P}}(1)$$

with $I(\theta_0) > 0$. Moreover, we postulate a six-time continuously differentiable, proper prior density π_λ . Then, using a Laplace approximation as in equation (2.6) of [2], we can claim that there is a random sequence $S_n = O_{\mathbb{P}}(1/\sqrt{n})$ not depending on the prior such that, for each $\lambda \in \Lambda$, the posterior mean satisfies

$$\left| \mathbb{E}_{\Pi_\lambda^{(n)}}(\theta) - \theta_0 - S_n \right| = \frac{1}{n} \left| \frac{\partial \log \pi_\lambda(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \frac{1}{I(\theta_0)} \right| + o_{\mathbb{P}}\left(\frac{1}{n}\right).$$

That is, up to a random term non depending on the hyperparameter, the gap between the posterior mean obtained with hyperparameter λ and the true parameter θ_0 depends on the magnitude of the first derivative of the log prior at θ_0 . Since this is null for each λ^* , it can be immediately seen that

$$\left| \mathbb{E}_{\Pi_{\lambda^*}^{(n)}}(\theta) - \theta_0 - S_n \right| = o_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Roughly, the discrepancies of posterior means from θ_0 driven by the prior choice are of lower asymptotic order for oracle selections of prior hyperparameters. If $\hat{\lambda}_n$ converges in probability to a point $\lambda^* \in \Lambda^*$ and the map $(\theta, \lambda) \mapsto (\partial/\partial \theta) \log \pi_\lambda(\theta)$ is continuous at (θ_0, λ^*) a similar conclusion holds under an empirical Bayes selection.

2.2 High-Dimensional Models

High-dimensional models offer a rather different picture compared to the previous one. In specific examples of parametric models with parameter dimension diverging along with the sample size, a notion of oracle hyperparameters choice is linked to priors (in a given class) yielding optimal posterior contraction rates. A notable example is the Gaussian sequence model with Horseshoe prior:

$$\begin{aligned} Y_i &= \theta_i + \varepsilon_i, & \varepsilon_i &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1), \\ \theta_i | \lambda, \tau_i &\sim \mathcal{N}(0, \lambda^2 \tau_i^2), & \tau_i &\sim \mathcal{C}^+(0, 1), \end{aligned} \tag{1}$$

for $i = 1, \dots, n$, where λ and τ_i are global and local shrinkage hyperparameters, while $\mathcal{C}^+$ is the half-Cauchy distribution. For each $\theta \in \mathbb{R}^n$, let $\|\theta\|_0 := \sum_{i=1}^n \mathbf{1}\{\theta_i \neq 0\}$. Then, with $p_n = \|\theta_0\|_0$ the number of truly nonnull $\theta_{0,i}$, postulate a sparse scenario where $p_n \rightarrow \infty$ but $p_n = o(n)$. In this setup, [7] prove that the oracle choice $\lambda^* = \frac{p_n}{n} \sqrt{\log(n/p_n)}$ leads to a posterior distribution which contracts at the minimax rate $\sqrt{p_n \log(n/p_n)}$ around the truth in L_2 distance. Moreover, Theorem 3.1-3.2 in [8] establish that the empirical Bayes procedure obtained by restricted MMLE

$$\hat{\lambda}_n = \operatorname{argmax}_{\lambda \in [1/n, 1]} \prod_{i=1}^n \int_{-\infty}^{\infty} \phi(Y_i - \theta_i) \left[\int_0^{\infty} \phi\left(\frac{\theta_i}{\lambda \tau_i}\right) \frac{1}{\lambda \tau_i} \frac{2}{\pi(1 + \tau_i^2)} d\tau_i \right] d\theta_i$$

yields a (pseudo) posterior distribution which concentrates near θ_0 at the same rate as the oracle one. A similar conclusion can be drawn for hierarchical posterior distributions ([8], Theorem 3.7), obtained by endowing λ with a prior supported on $[1/n, 1]$ and satisfying appropriate regularity conditions.

A similar phenomenon can be found for infinite-dimensional models. Consider for example the infinite-dimensional normal mean model

$$Y_i = \kappa_i \theta_i + \frac{1}{\sqrt{n}} Z_i, \quad Z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad \theta_i | \lambda \stackrel{ind}{\sim} \mathcal{N}(0, i^{-1-2\lambda}),$$

for $i = 1, 2, \dots$ where (Z_i) is a sequence of i.i.d. standard Gaussian random variables, (κ_i) is a sequence of known constants that decays to 0 as $i \rightarrow \infty$, precisely $C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p}$ for some $C \geq 1$ and $p > 0$, and $\theta = (\theta_i)$ is the parameter of interest, ranging over $\Theta = \ell_2$. If the true parameter sequence satisfies the summability property

$$\sum_{i=1}^{\infty} i^{2\lambda^*} \theta_{0,i}^2 < \infty,$$

then the oracle choice $\lambda = \lambda^*$ yields a posterior that cumulates its mass in ℓ_2 -balls centered at the true $(\theta_{0,i})$ with optimal rate $n^{-\lambda^*/(1+2\lambda^*+2p)}$. In contrast, a suboptimal posterior contraction rate is generally attained whenever $\lambda \neq \lambda^*$, see [3] Sect. 2. Optimal rates can be achieved also with restricted MMLE ([3], Theorem 2) equal to

$$\hat{\lambda}_n = \operatorname{argmax}_{\lambda \in [0, \log n]} \left\{ -\frac{1}{2} \sum_{i=1}^{\infty} \left(\log \left(1 + \frac{n}{i^{1+2\lambda} \kappa_i^{-2}} \right) - \frac{n^2}{i^{1+2\lambda} \kappa_i^{-2} + n} Y_i^2 \right) \right\},$$

and hierarchical posteriors ([3], Theorem 3) obtained with priors

$$\Pi_H(B) = \int_0^{\infty} h(\lambda) \Pi_{\lambda}(B) d\lambda,$$

provided that the hyperprior density h fulfills some mild assumptions.

A more general analysis is offered in [6] for nonparametric problems, where Θ with norm $\|\cdot\|$ is a Banach space. For each given $\lambda \in \Lambda$, assume the prior of $\|\theta - \theta_0\|$ is continuous under Π_{λ} and, for a given $K > 0$, also define the sequence

$\varepsilon_n(\lambda) \equiv \varepsilon_n(\lambda, M, \theta_0)$ via $\Pi_\lambda(\|\theta - \theta_0\| \leq K\varepsilon_n(\lambda)) \leq e^{-n\varepsilon_n(\lambda)^2}$. Under further regularity assumptions, it essentially represents the contraction rate at θ_0 of the posterior distribution originated from Π_λ . The best attainable rate with hyperparameter $\lambda \in \Lambda_n \subset \Lambda$ is then

$$\varepsilon_n^* = \inf \left\{ \varepsilon_n(\lambda) : \varepsilon_n(\lambda) \geq \sqrt{m_n \frac{\log n}{n}}, \quad \lambda \in \Lambda_n \right\}$$

with any sequence $m_n \rightarrow \infty$, provided that the infimum is taken on a nonempty set. The set of oracle priors yielding the best attainable posterior contraction rates is thus $\{\Pi_{\lambda^*}, \lambda^* \in \Lambda_n^*\}$, where $\Lambda_n^* := \{\lambda \in \Lambda_n : \varepsilon_n(\lambda) \leq M_n \varepsilon_n^*\}$ for any $M_n \rightarrow \infty$. Theorem 2.1 and Corollary 2.1 of [6] give conditions under which an empirical Bayes hyperparameter selection $\hat{\lambda}_n$ eventually lies in Λ_n^* and yields a (pseudo) posterior distribution which concentrates at the rate ε_n^* . Analogous results for contraction rates of hierarchical posteriors are provided in their Theorem 2.3.

3 Multiple Testing

Consider the Gaussian sequence model in (1), but this time equip the parameters θ_i with a spike and slab prior, i.e. with $\lambda \in (0, 1)$

$$\theta_i | \lambda \stackrel{i.i.d.}{\sim} (1 - \lambda)\delta_0 + \lambda G, \quad (2)$$

where δ_0 is the Dirac measure at 0 and G is a diffuse distribution on $\mathbb{R}$ whose density convolution with a standard Gaussian is g . In a large number of applications of model (1), it is very important to test the following multiple hypotheses $H_{0,i} : \theta_i = 0$ vs $H_{1,i} : \theta_i \neq 0$ simultaneously for $i = 1, \dots, n$. An intuitive way to do this when using a spike and slab prior is through a so-called ℓ -value procedure at a certain threshold t , i.e. with tests $\varphi_{\lambda,t} = (\mathbf{1}\{\ell_{i,\lambda} < t\})_{i=1}^n$, where $\ell_{i,\lambda} = \Pi_\lambda^{(n)}(\theta_i \neq 0)$ for $i = 1, \dots, n$. In this context, a natural risk measure combining false discovery rates and false negative rates of a procedure φ is $\mathcal{R}(\theta_0, \varphi) = \text{FDR}(\theta_0, \varphi) + \text{FNR}(\theta_0, \varphi)$. One may thus wonder which are the specific test statistics φ which yield the minimax risk $\inf_\varphi \sup_{\theta_0 \in \Theta_0} \mathcal{R}(\theta_0, \varphi)$, for a set of possible truths with s_n nonnull signals of minimum magnitude larger than a , i.e.

$$\Theta_0 \equiv \Theta_0(n) = \{\theta_0 : \|\theta_0\| = s_n, |\theta_{0,i}| > a \text{ for all } i \text{ s.t. } \theta_{0,i} \neq 0\},$$

for some $a > 0$ and (analogously to Sect. 2.2) with $s_n \rightarrow \infty$, $s_n = o(n)$.

If such boundary is set to $a = \sqrt{2 \log(n/s_n)} + b$, where b is a constant, and if a quasi-Cauchy density such that $g(y) = (2\pi)^{-1/2} y^{-2} (1 - \exp(-y^2/2))$ is chosen in prior specification (2), carefully scrutinising the proof of Theorem S-6 in [1] reveals that there are sequences of oracle values $\lambda^* \equiv \lambda^*(n) \asymp (s_n/n) \sqrt{\log(n/s_n)}$ such that at any fixed $t \in (0, 1)$ as $n \rightarrow \infty$

$$\sup_{\theta_0 \in \Theta_0} \mathcal{R}(\theta_0, \varphi_{\lambda^*,t}) = \inf_\varphi \sup_{\theta_0 \in \Theta_0} \mathcal{R}(\theta_0, \varphi) + o(1).$$

In words, ℓ -value procedures produced with oracle priors (asymptotically) attain the minimax risk. Theorem 2 in [1] establishes that the same is true for an empirical Bayesian procedure $\varphi_{\hat{\lambda}_n, t}$, obtained with

$$\hat{\lambda}_n = \operatorname{argmax}_{\lambda \in [1/n, 1]} \left\{ \sum_{i=1}^n \log \phi(Y_i) + \sum_{i=1}^n \log \left(1 + \lambda \left(\frac{g(Y_i)}{\phi(Y_i)} - 1 \right) \right) \right\}.$$

In fact, for any oracle sequence $\lambda^*(n)$, over a set sequence of probability tending to 1 it holds that $\hat{\lambda}_n \asymp \lambda^*(n)$; see Section S-8.2 of [1] for details.

References

1. Abraham, K., Castillo, I., Roquain, E.: Sharp multiple testing boundary for sparse sequences. [arXiv:2109.13601](https://arxiv.org/abs/2109.13601) (2023). <https://arxiv.org/abs/2109.13601>
2. Kass, R.E., Tierney, L., Kadane, J.B.: The validity of posterior expansions based on Laplace's method. In: Geisser, S., Hodges, J.S., Press, S.J., Zellner, A. (eds.) *Essays in Honor of George Bernard*, pp. 473–488. North Holland (1990)
3. Knapik, B.T., Szabó, B.T., van der Vaart, A.W., van Zanten, J.H.: Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Relat. Fields* **164**, 771–813 (2016)
4. Petrone, S., Rousseau, J., Scricciolo, S.: Bayes and empirical Bayes: do they merge? *Biometrika* **101**, 285–302 (2014)
5. Rizzelli, S., Rousseau, J., Petrone, S.: Empirical Bayes in Bayesian learning: understanding a common practice. [arXiv:2402.19036](https://arxiv.org/abs/2402.19036) (2024). <https://arxiv.org/abs/2402.19036>
6. Rousseau, J., Szabó, B.T.: Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Stat.* **45**, 833–865 (2017)
7. van der Pas, S.L., Kleijn, B.J.K., van der Vaart, A.W.: The horseshoe estimator: posterior concentration around nearly black vectors. *Electron. J. Stat.* **8**, 2585–2618 (2014)
8. van der Pas, S.L., Szabó, B., van der Vaart, A.W.: Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11**, 3196–3225 (2017)



Leveraging Pre-trained CNNs Feature Extractors for Classifying Genus *Oliva Bruguière*, 1789

Luca Romagnoli¹ (✉), Luigi Ippoliti², Luigi Di Carlo², and Cesare Brizio³

¹ University of Molise, Via F. De Sanctis, 86100 Campobasso, Italy
luca.romagnoli@unimol.it

² University G.d'Annunzio, Viale Pindaro 42, 65127 Pescara, Italy

³ World Biodiversity Association, Museo Civico di Storia Naturale di Verona, Lungadige Porta Vittoria, 9, 37129 Verona, Italy

Abstract. In this paper, we propose to evaluate the use of pre-trained convolutional neural networks (CNNs) as a features extractor to perform classification of the genus *Oliva Bruguière*, 1789. These mollusks include species whose overall smooth and ellipsoidal shape encompasses a limited degree of variability. The hallmark of this natural group, a combination of low interspecific variability and high intraspecific variability, has been for more than two centuries a source of taxonomic instability, including but not limited to the status of subgenera, or alleged genera, included in the *Oliva* clade. In order to assist the Linnean systematics of controversial groups, whose taxonomy is largely based on morphological landmarks, the present study explores the potential effectiveness of deep learning models as feature extractors for classifying the genus *Oliva*. In particular, by leveraging a dataset of about 1,000 digital images of *Oliva* shells categorized into 9 distinct subgenres, we employed pre-trained convolutional neural networks (CNNs) for feature extraction. Subsequently, applying statistical learning algorithms on a training set, we conducted a statistical classification analysis to assess the potential of an automated classification procedure in accurately predicting the labels of a test set. Preliminary results suggest that employing a pre-trained VGG16 architecture for feature extraction yields an accuracy of approximately 82% in classifying the test images.

Keywords: *Oliva Bruguière* 1789 · Pre-trained convolutional neural networks · Statistical learning · Image classification

1 Introduction

This paper outlines a preliminary approach for the application of automated image classification techniques to address the intricate challenges associated with a group of species that fall within the gastropod genus *Oliva Bruguière*, 1789. Unlike the conventional malacological approach, which primarily focuses on the study of the soft parts of shell-generating mollusks and is not amenable to the methods proposed herein, our investigation centers on a conchyliological and morphological framework.

The complexity inherent in the classification of genus *Oliva* arises from several key factors: 1) Low diversity in shell morphology: the shells of *Oliva* species exhibit an overall smooth, subovate/cylindrical/fusiform shape, lacking spines or prominent ornamentations that could facilitate both manual and automated recognition; 2) High intraspecific variability: within a single species, considerable variability in shell characteristics is observed, posing challenges for accurate classification and differentiation; 3) Low interspecific variability: conversely, there is minimal variation between species, complicating the task of distinguishing between different *Oliva* species based solely on morphological features; 4) Non-isometric growth: the growth patterns of *Oliva* species are not isometric, introducing additional complexity in the determination of species boundaries; 5) Wide range of adult axial sizes: *Oliva* species exhibit adult axial sizes spanning a considerable range, from 1 cm to 10 cm, creating difficulties in establishing a standardized framework for classification.

In light of these complexities, the proposed approach exploits digital morphology and automated image analysis techniques [1] to enhance the classification process.

2 Dataset and Pre-processing

A total of 948 pictures, representing *Oliva* shells belonging to 9 different subgenres, were considered for the statistical analysis. The distribution across classes reveals imbalances and results as follows (total specimens by subgenre): *Acutoliva* (47); *Americoliva* (190); *Annullatoliva* (75); *Carmione* (135); *Galeolella* (53); *Miniaceoliva* (170); *Musteloliva* (24); *Oliva* (141); *Viduoliva* (113).

To enhance the classification accuracy, a preprocessing step was implemented to isolate the *Oliva* shells by removing noise and background elements from the images. This allowed to focus the analysis solely on the shell morphology, thereby facilitating more accurate classification outcomes [2].

A visual representation of an example of pre-processed specimens are provided in Fig. 1.

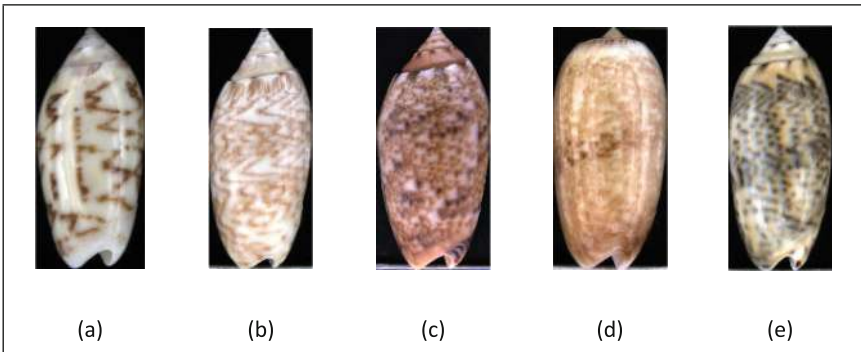


Fig. 1. Example of pre-processed images for some of the *Oliva* subgenres: (a) *Acutoliva*; (b) *Americoliva*; (c) *Annullatoliva*; (d) *Galeolella*; (e) *Miniaceoliva*.

3 Classification Results

As part of the preliminary statistical analysis, we advocate for the classification of Oliva shells utilizing pre-trained CNNs [3] as feature extractors. These models are able to transfer feature extraction knowledge into new models, and then train them with respect to a new specific classification task (Transfer Learning). Pre-trained models can be used in a wide range of computer vision tasks, including image classification, object detection, image clustering, and more. There are pre-trained models available for various architectures, among which VGG, ResNet, Inception, and MobileNet. Typically, the initial layers of CNNs capture fundamental input image attributes such as spots, boundaries, and color patterns, which are then refined by deeper hidden layers to construct more intricate higher-level feature patterns for improved image representation [5]. Each layer within the CNNs serves as an activation unit for the input images. Literature reviews suggest that when employing pre-trained CNNs for feature extraction, the features are commonly derived from the fully-connected layers immediately preceding the final output classification layer [4, 5].

For classification purposes, 717 images (approximately 75% of the total dataset) were randomly chosen as training set. Pre-trained CNNs from ImageNet, a vast database containing over 1.2 million images spanning 1,000 different classes, were then used for feature extraction. Specifically, the following three prominent deep CNN architectures were employed: 1) VGG16 (Vgg16); 2) ResNet50 version 2 (ResNet50_v2); 3) Inception-ResNet version 2 (IncResNet_v2).

The large number of features (Vgg16: 25,088 features; ResNet50_v2: 100,352; IncResNet_v2: 98,304) extracted from the algorithms right before the classification layer of the pre-trained CNNs, prompted us to reduce the input space for the classification step. As a second step of the analysis, we thus conducted one-way ANOVA tests on the features within each of the three feature matrices. The rationale behind the ANOVA tests lies in the fact that higher test values for a given feature indicate greater separation among the 9 subgenres. The outcome of this step was the reordering of features in ascending order of the p-values for each of the three matrices.

Based on the features selected through the ANOVA, we finally applied Linear Discriminant Analysis (LDA), k-nearest neighbors (KNN), and Support Vector Machine (SVM) for the final prediction task on 231 test images. Table 1 then presents the classification accuracies achieved by the three classifiers using different pre-trained CNNs.

As noticed, the optimal results consistently arise from the implementation of the Vgg16 architecture, regardless of the chosen classification algorithm. Further investigation is warranted for the utilization of ResNet50_2 and IncResNet_v2. Potential strategies for future research include:

- a) Exploring alternative methods for feature selection beyond ANOVA;
- b) Employing a more condensed ensemble of features by incorporating max pooling or average pooling downsampling operations during the feature extraction process;
- c) Identifying the optimal layer for feature extraction from pre-trained CNNs;
- d) Investigating transfer learning, fine-tuning, or customized CNNs tailored for classification purposes;

Table 1. Accuracy for 231 test set Oliva images for pre-trained CNN architectures and classification algorithms.

Pre-trained CNN architectures	Classification algorithms		
	LDA	KNN	SVM
Vgg16	81.8	78.4	81.0
ResNet50_v2	78.4	74.9	77.5
IncResNet_v2	68.4	55.8	64.9

- e) Implementing a multi-model ensemble strategy to combine predictions from multiple models;
- f) Leveraging the possibility of having dual views for each specimen (dorsal and ventral) to apply a multi-view classification model [6].

References

1. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53

2. Rother C., Kolmogorov V., Blake A.: “GrabCut”: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* **23**, 309–314 (2004)

3. Pintelas, E., Livieris, I.E., Kotsiantis, S., Pintelas, P.: A multi-view-CNN framework for deep representation learning in image classification. *Comput. Vis. Image Underst.* **232**, 103687 (2023)

4. Rajaraman, S., et al.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin bloodsmear images. *PeerJ*, 1–17 (2018)

5. Razavian, A.S., Azizpour, A., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519 (2014)

6. Seeland, M., Mäder, P.: Multi-view classification with convolutional neural networks. *PLoS ONE* **16**(1), e0245230 (2021). <https://doi.org/10.1371/journal.pone.0245230>



Explaining Item-Nonresponse in Open Questions with Requests for Voice Responses

Camilla Salvatore¹  and Jan Karem Höhne² 

¹ Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands
c.salvatore@uu.nl

² German Center for Higher Education Research and Science Studies (DZHW),
Leibniz University Hannover, Hannover, Germany
hoehne@dzhw.eu

Abstract. Conducting smartphone surveys offers flexibility in collecting various types of responses. Among these various response modalities, voice responses stand out for their potential to facilitate deeper respondent engagement and expression. However, high item-nonresponse rates pose significant challenges to their large-scale use. Therefore, the objective of this study is to better understand whether and to what extent socio-demographic characteristics, technological skills, and survey-related aspects are associated with item-nonresponse. Our preliminary findings suggest that certain socio-demographic groups, including males and respondents with low to medium levels of education, as well as younger respondents (aged 19 to 30 years), exhibit lower item-nonresponse rates. Additionally, respondents with good smartphone skills and those expressing high levels of interest in the survey show lower rates of item-nonresponse. This similarly applies to respondents perceiving the survey as being easy, not long, and not intimate.

Keywords: Smartphone surveys · item-nonresponse · open voice questions · respondent characteristics

1 Introduction

Survey agencies have been increasingly adopting web surveys due to their convenience and rapid data collection capabilities [1, 2]. A convenient way to complete web surveys is using a mobile device. Thus, the use of smartphones for web survey completion is increasing worldwide [3–5].

Smartphones introduce novel possibilities for web survey data collection, such as the acquisition of voice (or spoken) responses to open questions recorded through the built-in microphone. The use of voice input in web survey research is receiving increased attention, especially for its potential to gather detailed information that facilitate a deeper understanding of respondents' attitudes, beliefs, and behaviors [6].

Compared to traditional open text responses, open voice responses offer several advantages. Typing in text can be burdensome and time-consuming for respondents, requiring good literacy and writing skills. In contrast, voice responses require simply

pressing a recoding button, enabling respondents to freely express themselves without engaging in long information retrieval and text formulation processes [7]. Empirical studies comparing text and voice responses highlight the superior depth and richness of information gathered through voice responses [7, 8]. Schober et al. [9] show that voice responses result in notably higher data quality and increased reporting of sensitive information compared to traditional text responses. Similarly, Lütters et al. [10] show that voice responses result in more details, reduced completion times, and lower self-reported response burden than their text counterparts.

Despite the many advantages of requests for voice responses, there are some significant challenges. Voice responses are associated with comparatively high item-nonresponse rates: approx. 25% for voice responses to approx. 5% for text responses [6] and approx. 60% for voice responses to less than 5% for text responses [7]. In addition, Revilla and Couper [11] tested instructions explaining how to record voice responses in order to decrease item-nonresponse, but the authors did not find a decreasing effect. Item-nonresponse rates were still about 40%. Höhne [12] as well as Lenzner and Höhne [13] also show that respondents of younger age and higher survey interest are more willing to provide voice responses.

While existing efforts in the survey literature focus on investigating and improving instructions and technical aspects of voice responses [11], this study focuses on a better understanding of item-nonresponse when it comes to voice responses to open questions in smartphone surveys.

1.1 Research Questions

This paper is part of the initial phase of a comprehensive investigation of item-nonresponse in smartphone surveys with requests for voice responses to open questions. The objective of this study is to better understand whether and to what extent socio-demographic characteristics (e.g., age and gender), technological skills (e.g., smartphone skills and internet use), and survey-related aspects (e.g., interest and difficulty) are associated with item-nonresponse.

Through the examination of these factors, our research aims to provide novel insights into the determinants of item-nonresponse when it comes to voice responses. In doing so, we attempt to infer survey design recommendations and strategies for reducing missing data and improving data quality. However, survey respondents can exhibit different item-nonresponse patterns. We therefore distinguish between *voice skippers* (i.e. respondents who did not respond to any of the open voice questions but to the rest of closed questions) and *voice engagers* (i.e. respondents who responded to at least one of the open voice questions). We address the following two research questions (RQs):

RQ1. *Do voice skippers differ from the full sample with respect to socio-demographic characteristics, technological skills, and survey-related aspects?*

RQ2. *What socio-demographic characteristics, technological skills, and survey-related aspects drive item-nonresponse in open voice questions when it comes to voice engagers?*

To address these research questions, we leverage data obtained from an experimental study comparing open questions with requests for text and voice responses. In this paper, however, we only focus on voice responses because item-nonresponse poses a much

greater threat to voice than text responses (see [6, 7, 11]). A description of the methods and data is provided in the following section (Sect. 2). We then present our empirical findings in Sect. 3. Finally, in Sect. 4, we discuss the implications of our findings and outline avenues for future research.

2 Methods and Data

Data was collected in the Forsa Omninet Panel in Germany (November 2021). Forsa drew a cross-quota sample based on age and gender (3-by-2). They also drew three quotas on education. The email invitation included information on the device to be used (smartphone) and a survey link. The first page introduced the topic, outlined the procedure, and included a statement of confidentiality. Informed consent was obtained by Forsa and respondents received modest compensation for their participation.

In total, 1,022 respondents started the smartphone survey with open voice questions, of which 521 (or 51%) respondents dropped out. This leaves us with 501 respondents for analysis. These respondents were on average 48.7 (SD = 14.5) years old, and 48.3% of them were female. In terms of education, 31.2% had graduated from a lower secondary school (low education level), 42.5% from an intermediate secondary school (medium education level), and 26.3% from a college preparatory secondary school or university (high education level).

In this study, we consider eight open questions with requests for voice responses dealing with various topics, such as women at the workplace and media reports. Respondents could skip questions, but they were not provided with a non-substantive option, such as “don’t know.” The open voice questions were preceded by a voice response instruction. We used the open-source “SurveyVoice” [14] tool for recording voice responses. The tool resembles the voice input function of popular instant messengers, such as WhatsApp, and works on both iOS and Android smartphones.

3 Results

As this study serves as a preliminary investigation, we solely present descriptive results. More advanced analyses are planned for the full paper. Our analysis begins by examining the item-nonresponse profiles of respondents who participated in the smartphone survey. Table 1 shows the percentages of respondents by the number of questions they responded to. About 29.5% of the respondents did not provide a response to any of the eight open voice questions. Thus, they fall into the group of *voice skippers*. Conversely, about 57.8% of the respondents responded to all questions or missed just one.

Although respondents’ motives for opting not to respond to any voice questions remains unclear, there are two possible explanations: 1) respondents’ decision may stem from a change of mind regarding their willingness to respond through voice because it felt too intrusive or 2) some of the respondents may have faced technical difficulties in providing voice responses. Thus, these respondents were not willing and/or not able to respond. Next, to address the first research question (RQ1), we compare individual characteristics based on the item-nonresponse profile (*voice skippers* and *engagers*).

Table 1. Percentage of respondents who responded to a certain number of voice questions.

Number of questions responded to	None	1	2	3	4	5	6	7	All
% of respondents	29.5	1.2	0.4	0.8	1.4	2.6	6.4	21.0	36.8

Table 2 compares socio-demographic characteristics, technological skills, and survey-related aspects across the full sample, *voice skippers*, and *voice engagers*. *Voice skippers* tend to be slightly younger, male, and have lower interest in the survey than the full sample.

Since these respondents (*voice skippers*) seem to represent a special subgroup, we exclude them from the subsequent analyses and plan to conduct a more in-depth examination of this group in the full paper. We thus proceed with a detailed analysis of the *voice engagers* to address the second research question (RQ2). Table 3 shows the item-nonresponse rates by question and topic. On average, item-nonresponse rates are higher for questions associated with respondents’ emotions and feelings (Q1 and Q2) and probing questions associated with the understanding of the terms “global citizen” and “civil disobedience” (Q7 and Q8). Next, we study how item-nonresponse varies

Table 2. Comparison of sample characteristics across respondent groups.

Variables	Full sample	Voice skippers	Voice engagers
Age	48.7	47.1	49.3
Gender: Female	48.3	37.8	52.7
Education: Medium	42.5	41.6	42.8
Education: High	26.3	26.4	26.3
Smartphone skills	5.6	5.5	5.6
Internet use	6.0	6.0	6.0
Interest	5.4	4.6	5.7
Difficulty	3.1	3.3	3.0
Length	2.6	2.8	2.6
Intimacy	4.6	4.6	4.6
N	501	148	353

Note: We report means for age (in years), smartphone skills (1 “Very bad” to 7 “Very good”), internet use (1 “Not at all” to 7 “Very often”), survey interest (1 “Not interesting at all” to 7 “Very interesting”), survey difficulty (1 “Very easy” to 7 “Very difficult”), survey length (1 “Not long at all” to 7 “Very long”), and survey intimacy (1 “Not intimate at all” to 7 “Very intimate”). For gender (0 “Male” and 1 “Female”) and education (1 “Low,” 2 “Medium,” and 3 “High”), we report percentages instead.

across respondents' socio-demographic characteristics, technological skills, and opinions about the smartphone survey (survey-related aspects). For this purpose, we present some descriptive statistics of item-nonresponse rates in Fig. 1.

Focusing on socio-demographic characteristics, we note that item-nonresponse rates tend to be lower among males, respondents with low to medium levels of education, and those who are younger (aged 19 to 30 years). In relation to technological skills, respondents with at least good smartphone skills exhibit lower item-nonresponse rates, while there is no difference with respect to internet use. Regarding survey-related aspects respondents expressing high level of survey interest and evaluate the survey as being easy, not being long, and not being intimate, tend to produce lower item-nonresponse rates.

Table 3. Item-nonresponse rates (INR) for the voice engager group.

Question #	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
INR (%)	21.3	9.3	8.8	7.7	7.9	7.9	10.5	14.4
Topic	Emotion	Emotion	Refugees	Women	Media	Vaccine	Probe	Probe

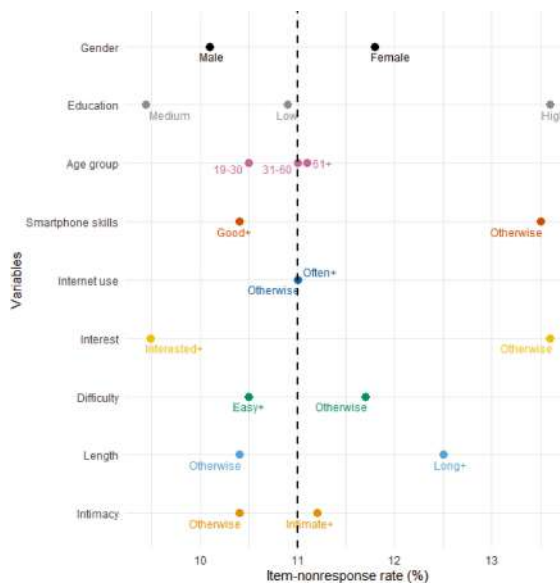


Fig. 1. Average item-nonresponse rates in the voice engager group. The dashed vertical line is the average item-nonresponse rate across all eight voice questions.

4 Conclusions

In conclusion, this study enhances our understanding of item-nonresponse when it comes to open questions with requests for voice responses, offering novel insights into this complex phenomenon. By distinguishing between *voice skippers* and *voice engagers*, we examined which socio-demographic characteristics, technological skills, and survey-related aspects vary across these two respondent groups. Moreover, we delved deeper into the factors associated with item-nonresponse among *voice engagers* showing that this group only slightly differs from the full sample. In the full paper, we aim to conduct further, more refined investigations to better understand the characteristics of *voice skippers* and *voice engagers*. We plan to carry out a more comprehensive assessment of item-nonresponse and its drivers including statistical tests in the form of multivariate analyses. This is accompanied by considering further variables, such as response length (or number of words), as a response quality measure. By continuing our investigations, we aim to provide valuable insights that provide empirical-driven strategies for mitigating item-nonresponse in open voice questions. The findings will be of interest for survey researchers and practitioners potentially contributing to an enhancement of data quality in future smartphone surveys.

References

1. Kumar, A., Chandran, S., Somani, A.: Conducting online web-based surveys at the time of COVID-19 pandemic: a short report. *Arch. Ment. Health* **22**(2), 158–161 (2021)
2. Couper, M.P.: The future of modes of data collection. *Public Opin. Q.* **75**(5), 889–908 (2011)
3. Peterson, G., Griffin, J., LaFrance, J., Li, J.: Smartphone participation in web surveys: choosing between the potential for coverage, nonresponse, and measurement error. In: *Total Survey Error in Practice*, pp. 203–233. John Wiley & Sons, Ltd. (2017)
4. Revilla, M., Toninelli, D., Ochoa, C., Loewe, G.: Do online access panels need to adapt surveys for mobile devices? *Internet Res.* **26**(5), 1209–1227 (2016)
5. Gummer, T., Höhne, J.K., Rettig, T., Roßmann, J., Kummerow, M.: Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. *Qual. Quant.* **57**, 5333–5353 (2023)
6. Gavras, K., Höhne, J.K., Blom, A.G., Schoen, H.: Innovating the collection of open-ended answers: the linguistic and content characteristics of written and oral answers to political attitude questions. *J. R. Stat. Soc. Ser. A Stat. Soc.* **185**(3), 872–890 (2022)
7. Revilla, M., Couper, M.P., Bosch, O.J., Asensio, M.: Testing the use of voice input in a smartphone web survey. *Soc. Sci. Comput. Rev.* **38**(2), 207–244 (2020)
8. Gavras, K., Höhne, J.K.: Evaluating political parties: criterion validity of open questions with requests for text and voice answers. *Int. J. Soc. Res. Methodol.* **25**(1), 135–141 (2022)
9. Schober, M.F., et al.: Precision and disclosure in text and voice interviews on smartphones. *PloS One* **10**(6), 1–20 (2015)
10. Luetters, H., Friedrich-Freksa, M., Egger, M.: Effects of speech assistance in online questionnaires, paper presented at the General Online Research Conference, Cologne, Germany (2018). <https://de.slideshare.net/luetters>
11. Revilla, M., Couper, M.P.: Improving the use of voice recording in a smartphone survey. *Soc. Sci. Comput. Rev.* **39**(6), 1159–1178 (2021)

12. Höhne, J.K.: Are respondents ready for audio and voice communication channels in online surveys? *Int. J. Soc. Res. Methodol.* **26**(3), 335–342 (2023)
13. Lenzner, T., Höhne, J.K.: Who is willing to use audio and voice inputs in smartphone surveys, and why? *Int. J. Mark. Res.* **64**(5), 594–610 (2022)
14. Höhne, J.K., Gavras, K., Qureshi, D.: *SurveyVoice (SVoice): A comprehensive guide for collecting voice answers in surveys* (2021)



IoT and AI for the Analysis of Emotional Intelligence

Vito Santarcangelo¹, Saverio Gianluca Crisafulli¹(✉), Giuseppe Oddo¹,
Gianfranco Piscopo², and Massimiliano Giacalone³

¹ iInformatica Srl, 75100 Matera, Italy
{vito,saverio}@iinformatica.it

² University of Naples “Federico II”, 80138 Napoli, Italy
gianfranco.piscopo@unina.it

³ University of Campania “L. Vanvitelli”, 81043 Capua, Italy
massimiliano.giacalone@unicampania.it

Abstract. This paper analyses the Internet of Things (IoT) and Artificial Intelligence (AI), two tools that can be applied for data analysis and collection, and whose processing and interpretation enables significant insights into emotional intelligence. In this regard, two application case studies that are emblematic of the integration of these technologies will be showcased, namely TokBek, an innovative smart intercom equipped with empathic design and artificial intelligence to monitor the quality of front-office service in offices, and VarGame, a VAR (Video Assistant Referee) system applied to table football that integrates Internet of Things artificial intelligence, demonstrating how these technologies can be used to collect relevant data and analyse emotional intelligence in different contexts.

Keywords: Internet of Things · Artificial Intelligence · Emotional Intelligence

1 Introduction

The digital age has ushered in an unprecedented technological revolution, in which concepts such as the Internet of Things and Artificial Intelligence have quickly established themselves as fundamental to progress in multiple sectors.

The Internet of Things is a concept that refers to the connection of physical devices to the Internet, allowing them to communicate with each other and with other systems, services or users [1]. The main goal of IoT is to make these devices intelligent, allowing them to collect data, transmit it, receive instructions and, in some cases, make decisions autonomously.

In parallel, AI has reached new heights, evolving from a science fiction vision to a tangible reality that permeates multiple aspects of our daily lives. AI encompasses a number of technologies that allow computer systems to emulate some human cognitive abilities, such as learning, reasoning and pattern recognition. This advanced processing capability allows AI to analyze large amounts of data efficiently and extract meaningful insights beyond human capabilities [2].

The union of IoT and AI has generated enormous analytical power, as IoT provides a vast amount of raw data and AI offers the tools to extract value from this data. In a world where the amount of data generated every day is constantly growing exponentially, the ability to efficiently collect, process and interpret this data has become a crucial competitive advantage for many industries and sectors.

The importance of these tools is further amplified by their role in data analysis and collection. IoT enables the creation of an ecosystem of interconnected devices that can collect real-time data from a variety of sources, including environmental sensors, wearables, and smart home devices. This data is then processed using AI algorithms, which can identify patterns, trends and relationships within it, providing valuable insights for optimizing operations, improving decisions and developing innovative solutions.

In this context, data analysis and collection extend well beyond mere numerical calculation, also embracing the analysis of human behaviors, emotions and social interactions. AI can be trained to understand and interpret emotional signals in the data collected, allowing emotional intelligence to be probed in ways previously unthinkable. This approach offers enormous potential to improve our understanding of human behavior and to develop innovative solutions.

The aim of this article is to provide an overview of these two tools, highlighting their potential in terms of data collection and analysis, including emotional intelligence, and then introducing two case studies implementing these technologies.

2 Internet of Things and Artificial Intelligence

The Internet of Things has roots that date back to the 1980s and 1990s, when the first attempts to connect physical objects to the Internet were made. However, it is in the last two decades that IoT has made significant progress, thanks to the advancement of sensor, networking and data processing technologies. The fundamental concept of IoT is to interconnect physical devices and everyday objects through sensors and internet connectivity, allowing them to collect data, communicate with each other and with external systems. Currently, IoT has reached an advanced state of development, with widespread adoption in multiple sectors such as health, agriculture, manufacturing, mobility, home automation and more. Millions of IoT devices are already in use and the number is expected to grow exponentially in the coming years. Key technologies involved in IoT include sensors, communication devices such as Bluetooth, Wi-Fi, NFC (Near Field Communication), cloud platforms for data storage and processing, and standard communication protocols such as MQTT, CoAP, and HTTP. IoT also often integrates with technologies such as Machine Learning and Edge Computing for real-time data analysis [3].

Artificial Intelligence has even older origins, dating back to the first experiments in the field of information technology and computational intelligence starting in the 1950s and 1960s. However, it is only in the last two decades that AI has made significant progress, thanks to the advancement of hardware, the accumulation of huge amounts of data, and the evolution of machine learning and deep learning algorithms. AI has reached an extraordinary level of development, with applications ranging from voice and image recognition to autonomous driving, machine translation, predictive analytics and

much more. Today, AI is ubiquitous in many aspects of our daily lives and continues to evolve rapidly. Key AI technologies include machine learning algorithms, deep learning, artificial neural networks, supervised and unsupervised learning algorithms, and data analysis tools [4].

These tools represent two pillars of the ongoing technological revolution, offering enormous opportunities for innovation and progress in multiple sectors. Their integration offers even greater potential, allowing you to collect data from a wide range of sources and use AI to extract meaningful insights and radically transform the way we interact with the world around us.

2.1 Data Analysis and Collection: Emotional Intelligence

The need to analyze and collect data arises from the growing amount of information generated by devices, sensors and interconnected systems, along with the need to gain meaningful insights from that data to support effective decisions, develop innovative solutions and improve user experience. Through IoT it is possible to create an ecosystem of interconnected devices capable of collecting real-time data from a wide range of sources, such as environmental sensors, wearable devices and smart home devices. This data can include information about temperature, humidity, geographic location, movement and much more. AI, on the other hand, offers the tools to analyze and interpret this data efficiently, identifying patterns, trends and relationships within it. Through the IoT, it is therefore possible to collect large quantities of data in real time from physical and virtual environments (for example, sensors mounted on industrial machinery can detect anomalies in production, wearable devices can monitor the physical activity and health of users, and smart home devices can record user behaviors within their homes), which are then processed using AI algorithms to extract meaningful insights (e.g. recognizing energy consumption patterns based on user behavior, identifying areas for improvement in supply chain or predict machinery failures based on anomalous signals detected by sensors) [5]. The benefits deriving from the joint use of IoT and AI in data analysis and collection are many. These include greater operational efficiency, better decision making, greater personalization of services, better resource management and greater innovation. However, one of the trends that is attracting the attention of many in recent times is the possibility of exploring and understanding more complex aspects of the human being, such as emotional intelligence.

Emotional intelligence is an essential component of human intelligence that concerns the awareness, understanding and management of emotions, both one's own and that of others [6]. IoT and AI can be used to analyze emotional intelligence by collecting data related to human emotions from a variety of sources, such as social interactions on social media, conversation recordings, biometric data and more (e.g. wearable sensors can monitor heart rate and body temperature to infer a person's emotional state, while AI algorithms can be trained to recognize facial expressions and tone of voice to assess an individual's level of stress or happiness).

Analyzing emotional intelligence through IoT and AI offers numerous practical benefits. It can be used to develop personalized emotional support systems that provide assistance and advice to people based on their emotional needs. In clinical settings, it can be used to monitor and manage emotional disorders such as anxiety and depression.

In a corporate context, it can be used to improve customer experience by personalizing services based on their emotional needs, and to improve staff management by assessing the emotional well-being of employees and acting accordingly.

3 Case Studies: TokBek and VarGame

As mentioned in the previous section, the Internet of Things and Artificial Intelligence are tools that can be applied for the analysis and collection of data, also enabling the analysis of emotional intelligence. For this reason, two application case studies integrating these technologies are presented in this paragraph.

TokBek is an innovative smart intercom, conceived by Angelo Cannarile of Sync Sas in synergy with the innovative SME and private research centre iInformatica, protected in design and by industrial invention patent application no. 102023000011214. The device features an empathic casket design, whose ergonomic shell is characterised by the presence of smart materials, and is equipped with an artificial intelligence module to monitor the quality of front-office service. The TokBek, which is equipped with multiple interfaces to connect to professional mixers (in recording studios) or to be used in front-office contexts (e.g. public counters of post offices/offices), also integrates data analysis functions that allow the analysis of operating times and modes, thus tracking the interactions between operator and external user for customer service purposes. Furthermore, thanks to the artificial intelligence module, it allows tone of voice and semantic analysis to estimate the quality and type of interaction, from an emotional intelligence perspective.

Vargame is an innovative system, designed and patented by Nicola Favale of Different Game Srl also in synergy with the innovative SME and private research centre iInformatica, which introduces the moviola concept into the world of table football. This is made possible by integrating IoT technology, thanks to the presence of sensors strategically distributed inside the goals, without altering the playing surface, whose operation is combined to detect goals in real time and suggest possible false positives/negatives to the VAR (referee's interconnected web dashboard). The match is also recorded by a dedicated camera that allows all the highlights to be shown on the spectators' smartphones, thus providing real-time feedback, and thanks to a computer vision-based artificial intelligence module, it is possible to track the positions and trajectories of the ball, thus allowing tactics and match dynamics to be analysed and evaluated/commented on. In this way, from the data collected via AI, it will be possible to provide real statistical data on the matches, as well as it will be possible to recognize game patterns, strategies and tendencies of the players, helping them to improve their skills and develop new tactics. Furthermore, from the data collected, it will also be possible to analyze the emotional intelligence of the various players. The system includes low energy consumption LEDs, positioned inside the table football table, specifically for sanitizing the ball in order to guarantee a hygienic playing environment (Fig. 1).

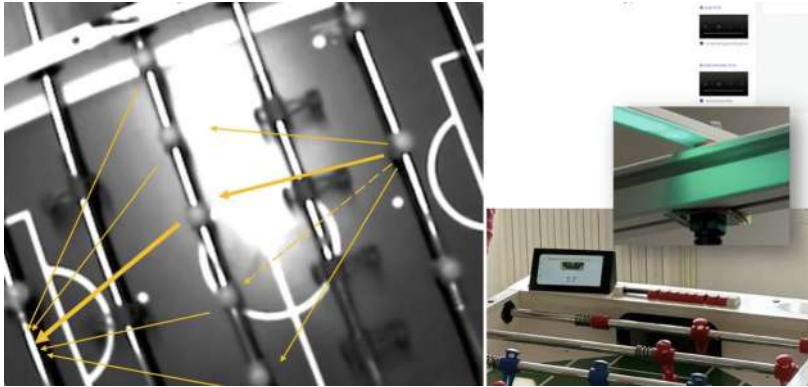


Fig. 1. The image on the left shows the trajectories of the ball analyzed by AI. In the image on the right you can see the dashboard, the sanitization LEDs and the result of the match in real time via smartphone (author's own).

4 Conclusions

In conclusion, this article has examined the crucial role of IoT and AI in data analysis and collection, with a particular focus on the growing importance of emotional intelligence. Through the integration of these two technologies, it is possible to collect and analyze data from a wide range of sources, allowing you to obtain meaningful insights to support effective decisions, develop innovative solutions and improve user experience. Furthermore, emotional intelligence analysis offers a new dimension in the use of IoT and AI, allowing to explore and understand complex aspects of the human being and to develop personalized solutions to improve people's emotional well-being and quality of life. In an era in which the volume of data continues to grow exponentially and interest in emotional intelligence increases, IoT and AI are confirmed as fundamental tools to face the challenges and seize the opportunities of the modern world.

References

1. Rose, K., Eldridge, S., Chapin, L.: The Internet of Things: an overview. In: The Internet Society (ISOC), vol. 80, pp. 1–50 (2015)
2. Ghosh, A., Chakraborty, D., Law, A.: Artificial intelligence in Internet of things. *CAAI Trans. Intell. Technol.* **3**(4), 208–218 (2018)
3. Shah, S.H., Yaqoob, I.: A survey: Internet of Things (IoT) technologies, applications and challenges. In: 2016 IEEE Smart Energy Grid Engineering (SEGE), pp. 381–385 (2016)
4. Allen, G.: Understanding AI technology. Joint Artificial Intelligence Center (JAIC) The Pentagon United States (2020)
5. Cui, Z., Jing, X., Zhao, P., Zhang, W., Chen, J.: A new subspace clustering strategy for AI-based data analysis in IoT system. *IEEE Internet Things J.* **8**(16), 12540–12549 (2021)
6. Salovey, P., Mayer, J.D.: Emotional intelligence. *Imagin. Cogn. Pers.* **9**(3), 185–211 (1990)



Leaders and Parties: How Do They Speak Online?

Francesco Santelli¹(✉), Alessia Forciniti², Valeria PolICASTRO³,
and Giancarlo Ragozini³

¹ DEAMS Department, Trieste University, Trieste, Italy
fsantelli@units.it

² Humanities Department, IULM University, Milan, Italy

³ Political Sciences Department, Naples University Federico II, Naples, Italy

Abstract. Social media usage has been increasing in political debates, but its role still needs to be defined. In this work, we want to investigate the differences in language usage between leaders' and parties' accounts in one of the most used social media platforms nowadays, Twitter. To deal with this aim, we defined a workflow to study different aspects of their differences using the data related to the 2022 Italian political campaign. Using other NLP and factorial approaches, this work aims to depict the overall online debate for this specific case, highlighting peculiar features for the type of account and political belonging.

Keywords: Twitter data · LDA · political debate topics · CAGALT · Italian politicians

1 Introduction

In modern electoral campaigns, the role of social media has become increasingly important; for instance, Twitter has become one of the most common “places” for political debate. Political leaders and parties increasingly used social media to spread information, and different researchers [2] started to study the relationship between online election campaigns and election results. People, as politicians, use social media in different ways; some use them more passively, and others in a more aggressive way (for example, Salvini in Italy [11] or Trump [12] in America). The role of social media in the political debate has always been discussed. Differently from works such as [5], which focuses on political-public relationships, in this paper, we would like to focus the attention on the differences in languages according to both a) political spectrum, i.e., left-right and b) individual leaders and political party accounts, so far never deeply highlighted for the Italian case in a quantitative approach, while descriptive and interesting works have been proposed especially for the 2018 electoral campaign [13]. To answer our objectives, different natural language processing techniques and factorial methods were applied to textual data collection to extract meaningful insights.

2 Online Textual Data: Collection and Preparation

The data refers to Italian political leaders involved in the last electoral campaign of 2022 and has been collected via API Twitter in the R software environment, using package `rtweet`. The selection of political actors was based on a prior qualitative analysis and scientific literature (e.g., [1]). Tweets have been retrieved from 2022-08-11 to 2022-09-20. Data extraction returned 19,949 tweets from 25 different accounts as shown in Fig. 1 from leaders and parties accounts, from left to right wing.

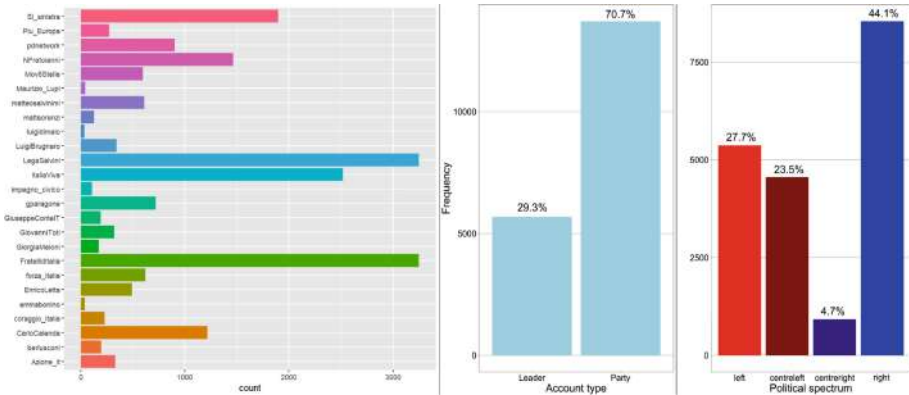


Fig. 1. Raw frequencies for the 25 accounts under analysis, and distribution of main metadata

The analysis of texts relies much on the choices related to preliminary data pre-processing, and no fixed procedures are established when analyzing textual data. Cleaning tools used can vary depending on the context, kind of corpora, research aims and even languages used. In the particular case of social media, peculiar attention should be devoted to the statistical analysis of texts, given the clear difference between this kind of corpora and traditional ones [7]. In this research case, data have been retrieved from Twitter (nowadays also called $\mathbb{X}$), and several features of this platform guided the textual preprocessing phase. First of all, the maximum length of each tweet is 280 characters (texts are pretty short). Language is, in over 99% of cases, pure Italian. Use of @ in the text has a specific meaning, that is to *mention* other users. Another symbol, #, is widely used to tag a Twitter status with the *hashtag*, helping users to search for that specific content. Thus, given the implicit information that they carry with them, we decided to include them in the analysis. Furthermore, in Italian is common that one single concept is expressed by 2 or even more words: for example, “reddito di cittadinanza” (*citizens’ income*), “campagna elettorale” (*electoral campaign*), “presidente della repubblica” (*republic president*) and so on. Taking into account all this preliminary contextual information, in this research we

adopted the following steps: 1. Created a corpus containing all the tweets and the metadata available (each row identifies a tweet). 2. Removed all the numbers and punctuation (except @ and #), given that in short messages punctuation is not so informative and tweets rarely consist of more than one sentence. 3. Removed stop-words for the Italian language, given that we are more interested in content defined by adjectives, nouns, and verbs. 4. Removed emoticons and URLs. In this phase, a few important emoticons carrying specific political meanings, have been included in the content 5. Text was transformed to lowercase (capital case was not so useful cause when dealing with proper names and accounts, usually symbol @ is included) 6. Bigram analysis: most common 25 included as one word (using “_” to link the two original words) 7. Additional stop-words found in the text have been removed manually, especially if having very high frequency From these steps, specific statistical methods adopted in this work for textual analysis will be described in Sect. 3.

3 Methods

3.1 LDA Analysis

For the purpose of this work, we performed an unsupervised Latent Dirichlet Allocation (LDA) [8] in order to discover, from the term-document matrix, the most important topics embedded in the corpus. For our tasks, we used an unsupervised technique because our aim is to extract relevant information in terms of subjects treated in this debate online and then use metadata (such as the account involved in the analysis) to deepen and characterize the findings. LDA is a cornerstone methodology in such a task, and also several extensions are nowadays available, even supervised versions, as in [9]. Its application to Natural Language Processing (NLP) is useful cause in a setting like LDA there is the aim of assigning words to topics and topics to documents in a generative way.

Furthermore, as proposed in Fig. 2, it is possible to visualize the similarity between topics as distances via Multidimensional scaling, using what is called the *intertopics distance* [10]. Using the first two dimensions of dimensionality reduction via Jensen-Shannon Divergence and Principal Components Analysis, topics are placed on the factorial map.

3.2 Factor Analysis

For what concerns factor analysis, we decided to adopt a technique that is able to handle both metadata available for each document (tweets) and the document-term matrix structure of the overall corpus. Due to this, we have chosen a CaGalt approach (Correspondence Analysis on Generalised Aggregated Lexical Table) as implemented in R package **FactoMineR** by [6]. This method allowed us to jointly analyze both the lexical table and political features in a multiple correspondence framework, combining in one larger table the matrix containing such features, and the document-term matrix. For CaGalt analysis, the X matrix is a block

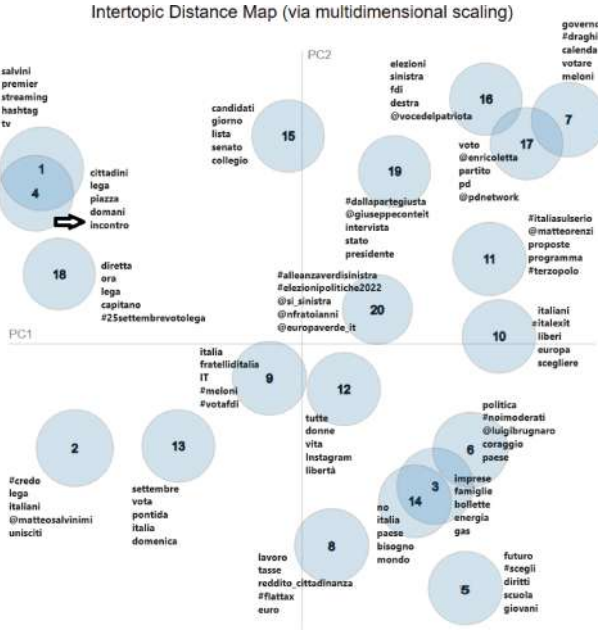


Fig. 2. Intertopic distance between 20 identified topics in LDA analysis. The first dimension splits the Lega and Fratelli d'Italia parties vs the others. The second dimension splits more programmatic content topics vs electoral propaganda topics.

matrix made up of two different matrices combined by columns: a DT document-term matrix and a V matrix with, for each document, the observed variables, as shown in the matrix representation below:

$$X_{n,(p+k)} = [DT_{n,p}|V_{n,k}] = \begin{bmatrix} dt_{1,1} & dt_{1,2} & \cdots & dt_{1,p} & v_{1,1} & \cdots & v_{1,k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ dt_{n,1} & dt_{n,2} & \cdots & dt_{n,p} & v_{n,1} & \cdots & v_{n,k} \end{bmatrix} \quad (1)$$

In this specific context, variables used as v , that is, metadata associated with the tweet, are three: account name, account type, and political spectrum.

4 Results

From the LDA analysis, the two main components highlight that the whole left-quadrant topics related to the Lega party and its leader, Matteo Salvini. Both accounts made massive use of Twitter during the campaign, and thus, this is reflected in the amount of topics identified related to this political area. In the upper part, the center and left parties and leaders tend to overlap, even with

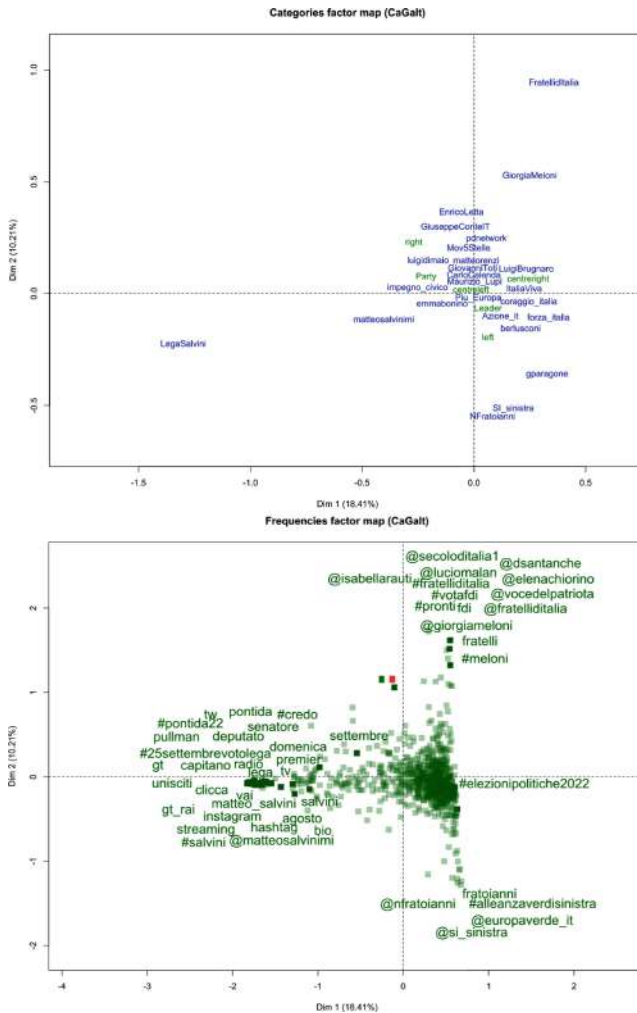


Fig. 3. Cagalt results. The first panel shows the account name in blue and the categories, which are account type and political spectrum, in green. Those are related to the second panel, showing the most important 50 words according to contribution to the factor analysis.

some divergences. While “Terzo polo” actors (liberal and centrist new coalition, guided by Matteo Renzi and Carlo Calenda) tend to debate more with other parties and leaders than to talk about themselves and thus use many terms that reflect such behavior. In the middle, two “extremes”: Sinistra Italiana (left) and Fratelli d’Italia (right). To some extent, their closeness means that they both tried to differentiate their identity from the liberal and democratic fields, without converging to some political agenda but rather using similar speech features. In

the bottom-right part, there are general topics, with no clear reference to any party, but with a slight predominance of central small parties.

For what concerns factorial analysis, on the first dimension we found leaders more related to the left part and parties to the right part, while on the second dimension, the political spectrum goes from right (upper part) to left (bottom part). Most of the time, parties and leaders are close to each other, with the exception of Carlo Calenda and its party Azione, and Matteo Salvini and its party Lega. In the middle, the generic hashtag *#elezionipolitiche2022*, widely used, stands out as the trending tag of this campaign (Fig. 3).

References

1. Loner, E.: Winners or losers? Using Twitter to evaluate how the COVID-19 pandemic changed the images of political leaders in Italy. *Contemp. Italian Polit.* **15**, 450–467 (2023)
2. Bracciale, R., Martella, A.: Define the populist political communication style: the case of Italian political leaders on Twitter. *Inf. Commun. Soc.* **20**, 1310–1329 (2017)
3. De Stefano, D., Santelli, F.: Combining sentiment analysis and social network analysis to explore twitter opinion spreading. In: *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, July 2019 (2019)
4. Carbone, F., Grasso, A., Santelli, F.: Populism and Negative Emotions within the Italian Politics: A Twitter-Based Analysis (2020)
5. Parmelee, J., Bichard, S.: *Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public* (Google eBook) (2011)
6. Kostov, B., Bécue Bertaut, M., Husson, F.: Correspondence analysis on generalised aggregated lexical tables (CA-GALT) in the FactoMineR package. *R J.* **7**, 109–117 (2015)
7. Rathore, A., Kar, A., Ilavarasan, P.: Social media analytics: literature review and directions for future research. *Decis. Anal.* **14**, 229–249 (2017)
8. Newman, D., Asuncion, A., Smyth, P., Welling, M.: Distributed algorithms for topic models. *J. Mach. Learn. Res.* **10** (2009)
9. Jelodar, H., et al.: Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools Appl.* **78**, 15169–15211 (2019)
10. Sievert, C., Shirley, K.: LDAvis: a method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70 (2014)
11. Berti, C., Loner, E.: Character assassination as a right-wing populist communication tactic on social media: the case of Matteo Salvini in Italy. *New Media Soc.* **25**, 2939–2960 (2023)
12. Collins, S., DeWitt, J.: Words matter: presidents Obama and Trump, Twitter, and US Soft power. *World Aff.* **186**, 530–571 (2023)
13. Bracciale, R., Cepernich, C., et al.: Hybrid 2018 campaigning: the social media habits of Italian political leaders and parties. *Italian Polit. Sci.* **13**, 36–50 (2018)



Meta-analytic Metagenomic Analysis of the Microbiome in Gastric Cancer

Gloria Santoro¹(✉), Elena Locatelli², Flavio Tirelli¹, Annamaria Agnes¹, Domenico D'Ugo¹, Roberto Persiani¹, Alberto Biondi¹, and Laura Lorenzon¹

¹ Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of the Sacred Heart, Rome, Italy

gloria.santoro@policlinicogemelli.it

² Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, TN, Italy

Abstract. Gastric cancer is a worldwide disease for which several risk factors are involved. *Helicobacter pylori* (*H. pylori*) infection is considered the main micro-organism impacting carcinogenesis. One of the new scientific approaches in gastric microbiome studies is the 16S ribosomal RNA (16S rRNA) amplification by liquid biopsies with next-generation sequencing (NGS) techniques. This study is a metagenomic meta-analytic investigation based on genomic data uploaded in online medical databases. These works were selected on the basis of the amplification of the 16S rRNA gene regions of the microbiomes in gastric biopsies aiming to describe the microbial taxonomy of the patients with *Helicobacter pylori* infections, distinguishing relative abundance in patients. Subgroups were differentiated according to pathological diagnosis, the molecular techniques used, the study design, and the geographical origin of the individuals.

1 Introduction

Gastric cancer is a worldwide disease and the hotspots of incidence and mortality exist in East Asia, Eastern Europe, and South America [1]. The stomach is divided in 5 parts: Cardia, Fundus, and Body (Corpus) make up the proximal stomach, and Antrum and Pylorus are part of the distal stomach. Based on topography, gastric cancer is defined as cardia e non-cardia depending on the anatomic site [2]. The tumor begins when cells in the stomach start to grow out of control due to different risk factors such as diet (high food intake of meat and salty meals with poor fresh fruits and vegetables diet), tobacco smoking, alcohol consumption and *Helicobacter pylori* (*H. pylori*) infection discovered in 1983 [3, 4].

H. pylori is considered the main micro-organism impacting carcinogenesis and it seems to trigger carcinogenesis. In addition, Epstein-Barr Virus (EBV), a human herpes virus, has been demonstrated a causal role in gastric carcinogenesis [5]. Despite the acid environment of the stomach, there are microorganisms able to survive within gastric cavity. On this basis, identifying a microbiome pattern in gastric cancer is a topic of scientific interest to define its population in disease and in healthy individuals, and to investigate if its modifications disrupt normal eubiosis [6].

Liquid biopsy is one of the new approaches developed to explore gastric microbiome. Samples are analyzed with Next Generation Sequencing (NGS) [7] amplifying 16S ribosomal RNA (rRNA) gene. This method has been widely used in the identification of bacterial sequences [8].

The prokaryotic rRNA genes comprehend the 5S, 16S, 23S, and intergenic regions [9]; the 16S rRNA gene is the most common marker gene for phylogenetic studies due to its presence in almost all bacteria [10]. The 16S rRNA gene includes nine hypervariable regions (V1-V9) interspersed between highly conserved regions [8, 11] used for Polymerase Chain Reaction (PCR) amplification of target sequences with universal primers [8]. The 16S hypervariable regions are used [12] to classify bacteria into different taxonomic groups, due to their sequence diversity among different species [8]. The 16S rRNA gene sequencing is carried out in many microbiome studies [11], but many different studies sequence only a part of this gene, ranging from one single variable region or spanning two/three regions (e.g., V3-V4, or V3-V5: ~300bp with Illumina sequencing platform) [9].

2 Aim of the Study

Using a metagenomic meta-analytic approach, this paper aims to investigate data collected from two different studies focusing on patients with gastritis, divided between individuals with and without *Helicobacter pylori* infection. Furthermore, it aims to compare microbial abundance and diversity between individuals analyzed in the East and in the West of the world.

3 Methods

The Human Microbiome Portal (HMP) and the National Center for Biotechnology Information (NCBI) BioProject section and Sequence Read Archive (SRA) were searched for eligible studies.

The following filters were applied to search literature and datasets: data involving both gastric cancer and microbiome data, using the following keywords: ((gastric cancer) AND (microbiota)) AND (microbiome); applying the following filters: Associated data, Clinical Study, Clinical Trial, Phase I, Clinical Trial, Phase II, Clinical Trial, Phase III, Clinical Trial, Phase IV, Dataset, Randomized Controlled Trial, Humans. The HMP was searched using: HMP data, Gastrointestinal Treat, 16S-GM-EA, Female and Male, Fastq, wgs_raw_seq_set.

Only papers and projects that performed the amplification of the 16S portion of the microbiota in their study design were considered, with attention given to selecting documents with data available in the bibliography allowing direct access to their 16S rRNA gene sequence data. Furthermore, the selection criteria primarily aimed to consider populations with a diagnosis of gastritis and categorization of the samples into the East and West of the world. We selected only files in Fastq format, but since Fastq files were not available in HMP, SRA identity codes were used to download complete data from the BioProject.

Two papers based on an Italian population cohort and a Chinese study analyzing only subgroups with the same gastritis diagnosis were considered. Moreover, both studies that performed the amplification of the 16S portion of the microbiota were included.

Bioinformatic data were downloaded via the bash command line (Ubuntu) by directly accessing the Bioproject online database through the SRA toolkit (<https://hpc.nih.gov/apps/sratoolkit.html>). The pre-processing analyses were re-evaluated using DADA2 (using software R version 4.3.1), and the data were subsequently analyzed by comparing them with the SILVA 132 reference database (<http://www.arb-silva>) containing aligned ribosomal RNA (rRNA) gene sequences from Bacteria, Archaea, and Eukaryota.

Alpha Diversity - Shannon Diversity Index (entropy) and Beta Diversity were explored. Alpha diversity measures the species diversity within samples, whereas beta diversity compares the diversity between samples. The distance between pairs of samples is calculated for individual studies. We calculated Bray–Curtis dissimilarity, quantifying the dissimilarity in species composition between two different sites, based on counts at each site matrix. We visualized results through a principal coordinate diagram (PCoA).

Finally, PERMANOVA constituted the final investigation to design a general linear model in which the independent variables of interest were evaluated.

4 Results

Our research led us to select scientific works and projects that span between 2010 and 2022. We focused the attention on two specific studies [13, 14] to review the results considering patients affected and not affected by Gastritis with some cases of *Helicobacter pylori* infection. The Italian research focused their attention on 24 patients diagnosed with gastritis and who were divided into 12 (50%) treated with proton pump inhibitors (PPI) and 12 (50%) untreated. Among the 12 patients treated with the drug, we have 8 patients (66.7%) negative for *H. pylori* infection and 4 (33.3%) positives, while in the group of untreated patients we have 7 patients (58.3%) negative for *H. pylori* infection, and the remaining 5 (41.7%) positive. We performed the microbiota analyses using a modern, updated and innovative approach, to define the relative abundance of the microbial population of patients diagnosed with gastritis. However, analyzing the data and considering the Phyla, Orders, Genus taxonomic levels, both comparing patients treated with PPIs and those patients untreated, do not present a different microbial population (Phylum, p-value: 0.18; Order, p-value: 0.06; Genus, p-value: 0.06). In the Chinese population, out of 410 patients analyzed, a subgroup of 104 patients had gastritis (54 *H. pylori* positive and 50 *H. pylori* negative). The same analyses were carried out for the Phyla, Orders, Genus taxonomic levels, documenting a statistically significant difference between the microbiologic population of *H. pylori* positive vs negative patients (Phylum, p-value: <0.01; Order, p-value: <0.01; Genus, p-value: <0.01). This difference in results suggests that, in addition to considering different analysis cohorts and having different assumptions in the study design, an important discrepancy arises from the molecular techniques employed. Specifically, the NGS techniques used were 454 pyrosequencing for the Italian cohort, while Illumina MiSeq machinery was utilized for the Chinese cohort.

5 Conclusion

Meta-analytic metagenomic analysis enabled us to attain a more comprehensive understanding of the current state of research on the gastric microbiome. However, due to the significant sequence diversity resulting from various available NGS platforms (such as 454 pyrosequencing and Illumina MiSeq), a single hypervariable region cannot fully elucidate and distinguish among all bacteria [8]. Furthermore, the technologies utilized exhibit notable differences in their ability to identify novel bacterial sequences. While 454 pyrosequencing technology offers longer read sequences and greater flexibility, making it better suited for gene sequences with higher GC content, it is also a more expensive and slower technology for laboratory operations, and ultimately less precise than Illumina MiSeq. These characteristics may better account for the discordant results observed despite employing the same statistical techniques.

Moreover, the reference database used (SILVA) in these recent analyses is more updated, implying advancements in microbiological knowledge that may not yield statistically significant differences at the taxonomic level among Phyla, Orders, and Genus. Delving deeper into taxonomic levels, such as considering Species, might have revealed potential differences. However, such analyses would require a substantial amount of computational power and computers with greater memory.

In conclusion, a more appropriate choice of sample size in the Italian study design could have led to increased reproducibility and more precise analyses, even in light of recent discoveries in the international microbial field. For this reason, new NGS technologies are being developed to facilitate deeper analysis of the microbiome genome. Thus, this meta-analytic work has served to summarize the current understanding of the gastric microbiota while also underscoring the need for further in-depth analysis.

References

1. Smyth, E.C., Nilsson, M., Grabsch, H.I., van Grieken, N.C., Lordick, F.: Gastric cancer. *Lancet* **396**, 635–648 (2020). [https://doi.org/10.1016/S0140-6736\(20\)31288-5](https://doi.org/10.1016/S0140-6736(20)31288-5)
2. Rawla, P., Barsouk, A.: Epidemiology of gastric cancer: global trends, risk factors and prevention. *Prz. Gastroenterol.* **14**, 26–38 (2019). <https://doi.org/10.5114/PG.2018.80001>
3. Ianiro, G., Molina-Infante, J., Gasbarrini, A.: Gastric microbiota. *Helicobacter* **20**(Suppl 1), 68–71 (2015). <https://doi.org/10.1111/HEL.12260>
4. Marshall, B.J., Warren, J.R.: Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311–1315 (1984). [https://doi.org/10.1016/S0140-6736\(84\)91816-6](https://doi.org/10.1016/S0140-6736(84)91816-6)
5. Sitarz, R., Skierucha, M., Mielko, J., Offerhaus, G.J.A., Maciejewski, R., Polkowski, W.P.: Gastric cancer: epidemiology, prevention, classification, and treatment. *Cancer Manag. Res.* **10**, 239–248 (2018). <https://doi.org/10.2147/CMAR.S149619>
6. Barra, W.F., et al.: Gastric cancer microbiome. *Pathobiology* **88**, 156–169 (2021). <https://doi.org/10.1159/000512833>
7. Goodwin, S., McPherson, J.D., McCombie, W.R.: Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**(6), 333–351 (2016)
8. Chakravorty, S., Helb, D., Burday, M., Connell, N., Alland, D.: A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* **69**, 330–339 (2007). <https://doi.org/10.1016/J.MIMET.2007.02.005>

9. Church, D.L., Cerutti, L., Gürtler, A., Griener, T., Zelazny, A., Emler, S.: Performance and application of 16S rRNA gene cycle sequencing for routine identification of bacteria in the clinical microbiology laboratory. *Clin. Microbiol. Rev.* **33**, 1–74 (2020). <https://doi.org/10.1128/CMR.00053-19>
10. Janda, J.M., Abbott, S.L.: 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764 (2007). <https://doi.org/10.1128/JCM.01228-07>
11. Johnson, J.S., et al.: Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10** (2019). <https://doi.org/10.1038/S41467-019-13036-1>
12. Hassler, H.B., et al.: Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. *Microbiome* **10** (2022). <https://doi.org/10.1186/S40168-022-01295-Y>
13. Cong, H.E., et al.: Convergent dysbiosis of gastric mucosa and fluid microbiome during stomach carcinogenesis. *Gastric Cancer* **25**(5), 837–849 (2022)
14. Paroni Sterbini, F., et al.: Effects of proton pump inhibitors on the gastric mucosa-associated microbiota in dyspeptic patients. *Appl. Environ. Microbiol.* **82**(22), 6633–6644 (2016)



Estimating a Quantile-Based Income Inequality Indicator on Complex Sampling Data

Silvia Scarpa^{1,2(✉)}, Maria Rosaria Ferrante², and Stefan Sperlich³

¹ Department of Statistical Sciences, University of Padova, Padova, Italy
silvia.scarpa.6@phd.unipd.it

² Department of Statistical Sciences, University of Bologna, Bologna, Italy

³ Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland

Abstract. Several statistical indicators exist for the measurement of economic inequality. They are mostly based on distribution moments or depend on few percentiles at the tails of the distribution, selected a priori. This work analyzes a recently proposed quantile-based income inequality indicator, the Quantile Ratio Index, which solely depends on quantiles and considers the whole distribution. A complete finite population estimation framework for this indicator is here proposed, that works for data collected with complex sampling design. In order to obtain a reliable measure, special attention is dedicated to the task of estimating quantiles, particularly when accounting for sampling weights. Simulations based on Italian EU-SILC 2017 data demonstrates that the proposed direct estimator has large accuracy, precision and robustness to outlying observations, if based on an appropriate choice of quantile estimator.

Keywords: EU-SILC data · income inequality · quantile estimation · quantile ratio index

1 Introduction

The Gini coefficient is the most widely used tool to measure income inequality. However, it has some limitations, such as its sensitivity to outliers in the upper tail of the income distribution. The literature offers a plethora of alternative indicators and, according to the research interest, one indicator may be more appropriate than another. The recently proposed *Quantile Ratio Index* by [5], QRI henceforth, gained our attention for different reasons. It is solely based on quantiles, as it exploits their robustness properties, having limited sensitivity to anomalous values. [5] introduced the QRI together with an estimator. Their definitions are based on distributions referring to infinite (hyper-)populations. The present work aims at providing a complete methodology that works for both, infinite (hyper-) and finite populations. An estimator of the QRI, with the inclusion of sampling weights, and of its sampling variance are proposed. In order to do so, we first tackle the challenge of estimating quantiles in finite populations.

2 The Quantile Ratio Index

The QRI, originally proposed by [5], is an income inequality indicator based on the ratio between symmetric quantiles. Let Y be a random variable denoting positive income with absolutely continuous cumulative distribution function F , which admits quantile function $Q(p)$. Define for any $0 < p < 1$ the ratio between symmetric quantiles as $R(p) = \frac{Q(p/2)}{Q(1-p/2)}$. The QRI is defined as

$$\text{QRI} = \int_0^1 (1 - R(p)) dp = \int_0^1 \left(1 - \frac{Q(p/2)}{Q(1-p/2)} \right) dp, \quad 0 \leq \text{QRI} \leq 1. \quad (1)$$

The QRI measures the distance between the income equi-distribution situation and the average ratio among the incomes below the median and their symmetric counterparts above the median.

Consider now a finite population U of N identifiable units, with value y_i being the positive income of unit i of population U . Consider then a grid of M cells, each defined on $(0, 1)$ as $p_m = \frac{m-1/2}{M}$, $m = 1, \dots, M$. The finite population quantiles can be computed on each grid cell as $Q_N(p_m)$. The ratio between symmetric quantiles is evaluated on the grid by $R_N(p)$. Then the QRI in the finite population U is defined as

$$\text{QRI} = \frac{1}{M} \sum_{m=1}^M (1 - R_N(p_m)) = \frac{1}{M} \sum_{m=1}^M \left(1 - \frac{Q_N(p_m/2)}{Q_N(1-p_m/2)} \right). \quad (2)$$

2.1 Estimation of the QRI and the Quantiles in Finite Population

When dealing with socio-economics aspects, the full population data are rarely available. Instead, survey data are made accessible to researchers. Therefore, a reliable estimator of the parameter of interest is thus necessary. Imagine to draw a random sample S of size n from U by means of a complex sampling design. Let w_j be the sampling weight of unit j in the sample, that can be calibrated on known marginal totals, or adjusted for non response. Let $W_j = \sum_{i \in S} w_i \mathbb{1}(i \leq j)$ denote the cumulative sum of the weights up to observation j . Consider as before a grid approximation in $(0, 1)$ of a sufficiently large number of cells, M . Let $\hat{Q}(p_{m/2})$ be the $p_{m/2}$ -th finite population quantile estimator. Then the QRI estimator is given by

$$\widehat{\text{QRI}} = \frac{1}{M} \sum_{m=1}^M \left(1 - \frac{\hat{Q}(p_{m/2})}{\hat{Q}(1-p_{m/2})} \right), \quad (3)$$

$\widehat{\text{QRI}}$ strongly depends on the selection of a proper quantile estimator. There exist many definitions in literature which provide nonparametric estimators of their population counterpart when sampling weights are omitted. This work uses the following general rule for the estimation of the quantiles in the finite population framework:

$$\hat{Q}(p) = Y_{(k-1)} + (Y_{(k)} - Y_{(k-1)}) \left(\frac{p - \hat{r}_{k-1}}{\hat{r}_k - \hat{r}_{k-1}} \right) \quad (4)$$

where $\hat{r}_k$, the plotting position, indicates the estimator of the cdf and the selected k is such that $W_{k-1} - m_{k-1} < W_n p < W_k - m_k$. According to the definition of r_k and m_k , a different quantile estimator is obtained. We develop here extensions of the continuous linear interpolation rules reviewed in [3] to complex sampling design data, in order to include the sampling weights. Table 1 displays the considered estimators of the cdf, and the corresponding interpolation parameters. The quantile estimators are indexed with the same enumeration from four to nine as the one used in the paper by [3].

Table 1. Extension of [3] plotting positions to the incorporation of the sampling weights

Rule	$\hat{r}_k$	$\hat{m}_k$
$\hat{Q}_4(p)$	$\frac{W_k}{W_n}$	0
$\hat{Q}_5(p)$	$\frac{W_k - \frac{1}{2}w_k}{W_n}$	$\frac{w_k}{2}$
$\hat{Q}_6(p)$	$\frac{W_k}{W_n + w_n}$	$w_n p$
$\hat{Q}_7(p)$	$\frac{W_{k-1}}{W_n - 1}$	$w_k - w_n p$
$\hat{Q}_8(p)$	$\frac{W_k - \frac{1}{3}w_k}{W_n + \frac{w_n}{3}}$	$\frac{w_k}{3} + \frac{w_n}{3}p$
$\hat{Q}_9(p)$	$\frac{W_k - \frac{3}{8}w_k}{W_n + \frac{1}{4}w_n}$	$\frac{3}{8}w_k + \frac{w_n}{4}p$

This work considers also an extension of the [2] quantile estimator to complex survey data, that includes sampling weights ($\hat{Q}_{HD}(p)$) and the Eurostat recommended quantile estimator ($\hat{Q}_{Laek}(p)$).

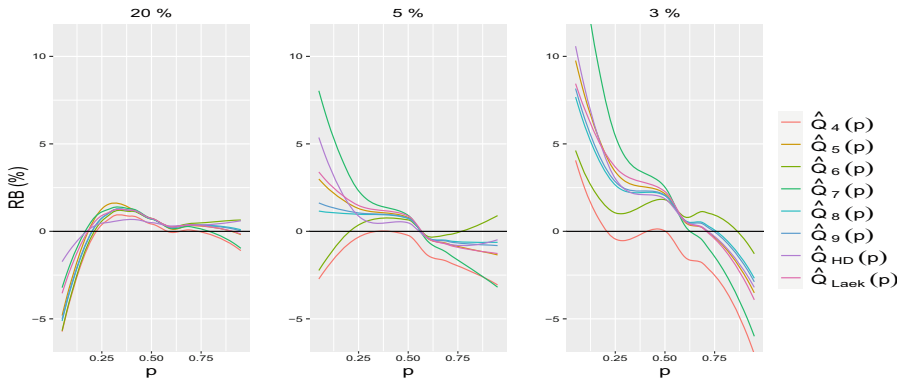


Fig. 1. RB (%) against p by quantile estimators on simulated samples from EU-SILC data (Abruzzo).

Simulations can demonstrate that the quantile estimators have different behaviours, in terms of bias and MSE, especially when small sample sizes and observations at the tails of the distribution are involved. Figure 1 shows the relative bias in percentage, for

any order $0 < p < 1$, of the considered quantile estimators, when 1000 samples are drawn from the Italian EU-SILC 2017 data. Draws are done with three sampling rates (20%, 5%, 3%), replicating the original sampling scheme, that is a two-phase stratified design, and considering the Italian regions as strata variables. Figure 1 reports the results for Abruzzo region, with average sample size respectively of 213, 72 and 29.

Depending on the selected l -th quantile estimator among the one just introduced, a different QRI estimator, $\widehat{QRI}_l$, is obtained. The same simulation as before is used to test the statistical behaviour of the eight here proposed QRI estimators, and to compare them with estimators of already known inequality indicators, such as the Gini coefficient ($\widehat{G}$), the Quintile Share Ratio ($\widehat{QSR}$), the Palma index ($\widehat{Palma}$) and the P90/P10 ratio ($\widehat{P90/P10}$). Except for the first one, they all depend on percentiles. For these estimators, the Eurostat quantile estimator is used.

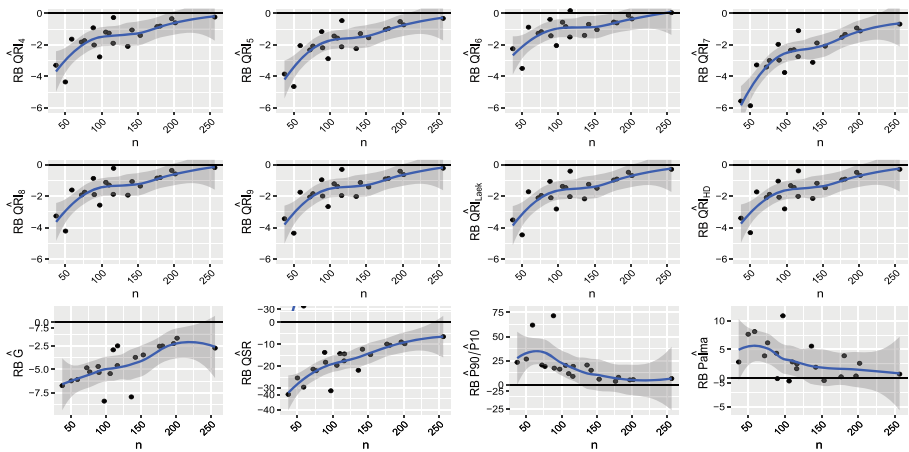


Fig. 2. Relative bias against regions sample size, with sampling rate 5%.

In Figure 2, the relative bias (RB), in percentage, is used as measures of performance. The first eight plots show that the QRI estimators behave differently. Particularly, $\widehat{QRI}_6$ shows the smallest RB. Moreover, the QRI estimators demonstrate less variability and less bias than the estimators of the other inequality indicators, even when the sampling size is low. Similar results are obtained if model-based simulations are applied.

2.2 Distribution and Sampling Variance of the QRI Estimators

The $\widehat{QRI}_l$ distribution, for any considered l , resulting from the performed simulations is approximately Normal as long as $n > 30$, though, certainly, with bounded support in $(0, 1)$.

The sampling variance can be estimated using an adapted rescaled method, the latter going back to [6], that works for a general stratified multistage sample. Simulations

can demonstrate that this method works better than naive bootstrap or superpopulation resampling methods. Linearization methods based on the influence function do not work in this case, due to the fact that the influence function of the QRI is discontinuous and requires the estimation of both, the quantiles and of the density function. Nevertheless, the influence function, derived using the differential rules as in [4], is bounded, ensuring that extreme values or a contamination do not have a strong influence on our inequality indicator [1]. On the opposite, moment-based indicators, such as the Gini coefficient, have unbounded influence functions. The QRI demonstrates to be robust to outlying recorded observations, and also to have a small standard error, ensuring reliability for statistical inference.

3 Concluding Remarks

This works shows that the proposed QRI estimator for data coming from complex sampling surveys, based on an accurate choice of the quantile estimator, has smaller bias and variability than other common inequality estimators, giving much smaller mean squared errors. The QRI estimator sampling variance can be estimated using rescaled bootstrap methods, that consider the original survey sampling weights and preserve the sampling scheme information. The QRI influence function is bounded, ensuring robustness to outlying values and limited bias, even in presence of very large contamination. These characteristics ensure that the QRI is a reliable statistical tool for the measurement of inequality.

References

1. Cowell, F.A., Victoria-Feser, M.-P.: Robustness properties of inequality measures. *Econometrica* **64**(1), 77–101 (1996)
2. Harrell, F.E., Davis, C.E.: A new distribution-free quantile estimator. *Biometrika* **69**(3), 635–640 (1982)
3. Hyndman, R.J., Fan, Y.: Sample quantiles in statistical packages. *Am. Stat* **50**(4), 361–365 (1996)
4. Osier, G.: Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Surv. Res. Methods* **3**(3), 167–195 (2009)
5. Prendergast, L.A., Staudte, R.G.: A simple and effective inequality measure. *Am. Stat* **72**(4), 328–343 (2018)
6. Rao, J., Wu, C., Yue, K.: Some recent work on resampling methods for complex surveys. *Surv. Methodol.* **18**(2), 209–217 (1992)



Exploring New Perspectives on the Ukrainian-Russian Conflict Using Text Mining Techniques: An Updated Analysis

Leonardo Scarso^{1(✉)}, Marco Novelli², and Francesco Saverio Violante¹

¹ Occupational Medicine Unit, Department of Medical and Surgical Sciences,
University of Bologna, Via Pelagio Palagi 9, 40138 Bologna, Italy
{leonardo.scarso2, francesco.violante}@unibo.it

² Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41,
40126 Bologna, Italy
m.novelli@unibo.it

Abstract. Within the context of contemporary global conflicts, this paper undertakes an extensive examination, focusing notably on the ongoing Russia-Ukraine dispute. Drawing from a wide array of scholarly research, we embark on a nuanced exploration of the intricate layers characterizing these conflicts, ranging from geopolitical complexities to socio-economic repercussions. Through the application of advanced clustering methodologies, we discern and categorize distinct thematic clusters, highlighting key issues such as the role of social media dynamics, the economic implications, and the diplomatic intricacies. Our analysis offers practical insights for policymakers and stakeholders grappling with the intricate challenges of conflict resolution. By expanding the analytical scope and updating the dataset with recent publications, this study provides a comprehensive understanding of the contemporary conflict and its enduring ramifications.

Keywords: Model-based clustering · cosine distance · mixture models · textual data

1 Introduction

In recent years, scholarly investigations have been exploited for the exploration of the intricacies of geopolitical tensions, providing valuable perspectives on the origins, dynamics, and consequences of global conflicts. These scholarly researches not only enhance our comprehension of particular disputes, but also can treat broader issues encompassing diplomacy, security, and human rights. The ongoing study represents an extension of a prior investigation centered on the analysis of titles and abstracts related to the conflict between Russia and Ukraine [1]. Building upon the findings of the earlier research, which depicted seven distinct research paper topics published from February 24th, 2022, to April 15th, 2023, the current research expands its scope.

The dataset of papers has been updated as of December 31st, 2023. This addition enriches the dataset, enabling a more comprehensive exploration of contemporary conflicts and their far-reaching consequences beyond the immediate war zones. The primary aim is to discern and categorize the key subjects discussed by scholars regarding the persisting conflict. Specifically, this research will provide a detailed update on Ukrainian-Russian themes to ascertain whether the trajectory of topics from the preceding study has been sustained or if new issues have emerged.

2 Methodology

In analyzing abstracts on the Ukraine-Russia conflict, traditional natural language processing and machine learning methods may fall short due to sparse data and lack of prior knowledge. To address these critical issues, a model-based clustering technique using cosine-based distance densities, as proposed in [2], is adopted. This method aims to create a ‘taxonomy’ of statistical papers by leveraging the high sparsity of the document-term matrix, wherein every cell denotes the frequency of a term’s occurrence within an abstract, with sparsity indicating the percentage of zeros in the matrix. Cosine distance is chosen for its reliability in measuring dissimilarity between documents, particularly in sparse matrices, focusing solely on non-zero entries. The proposal involves employing a mixture of k cosine distance-based density functions to effectively categorize the abstracts:

$$f(\mathbf{y}; \boldsymbol{\xi}, \lambda) = \sum_{i=1}^k \pi_i \psi(\lambda) e^{-\lambda d(\mathbf{y}, \boldsymbol{\xi}_i)},$$

where:

- π_i , $i = 1, \dots, k$ are the positive mixture weights, such that $\sum_{i=1}^k \pi_i = 1$;
- $\boldsymbol{\xi}_i$, $i = 1, \dots, k$ are the component varying centroid vectors;
- $d(\mathbf{y}, \boldsymbol{\xi}_i)$ is the cosine distance between the p -dimensional document $\mathbf{y}$ and the respective centroid $\boldsymbol{\xi}_i$, that is:

$$d(\mathbf{y}, \boldsymbol{\xi}_i) = 1 - \frac{\sum_{h=1}^p y_h \xi_{ih}}{\sqrt{\sum_{h=1}^p y_h^2} \sqrt{\sum_{h=1}^p \xi_{ih}^2}},$$

- with y_h and ξ_{ih} denoting the frequency of word h in $\mathbf{y}$ and $\boldsymbol{\xi}_i$, respectively;
- $\psi(\lambda)$ is the normalizing constant, depending on the precision parameter λ , with $\lambda > 0$ acting as a scaling factor of $d(\mathbf{y}, \boldsymbol{\xi}_i)$.

Parameters estimation is performed via EM algorithm, initialized with the solution of the spherical k -means based on the cosine distance [3].

3 Data Analysis

In this part of the article, we will delve into the conducted examination and the methodology employed, starting with a short explanation of the collected information.

3.1 Data Set Description

The dataset contains 2,745 entries comprising scientific articles published between February 24th, 2022, and December 31st, 2023, on ongoing conflicts, extracted with a query containing the words ‘war’, ‘Russia’, ‘Ukraine’, ‘conflict’. It includes article title, main authors, DOI, type, publisher, year, and abstract. Data were sourced from Google Scholar (2,504 articles) and PubMed (241 articles). Most articles were from 2023 (58.3%), mainly journal articles and book chapters (1,970 and 436, respectively). Article titles were gathered with abstracts. Pre-processing involved removing special characters, punctuation, common words, and stemming. Excluding query terms, three-character tokens, numeric terms, and infrequent words, alongside removal of terms with less than 20 occurrences, led to a decrease in sparsity to 93.1% from the initial 99%.

3.2 Results and Discussion

To find the ideal number k of clusters representing the distribution of scientific abstracts on conflicts, the EM algorithm is run with varying k values from 2 to 20. The Bayesian information criterion (BIC) was used to determine the optimal number of clusters. Cohesion index C_i gauges cluster homogeneity, calculated via cosine distance metric:

$$C_i = \sqrt{1 - \bar{d}_i^2}, \quad 0 \leq C_i \leq 1,$$

where, $\bar{d}_i$ is the average cosine distance within cluster i . A value of C_i close to 1 denotes greater intra-cluster homogeneity. Clusters quality is also assessed via Average Silhouette Width (ASW) [4], gauging cluster separation and appropriateness of assignments. For each observation, silhouette width s_j is computed as:

$$s_j = \frac{b_j - a_j}{\max(a_j, b_j)}, \quad j = 1, \dots, n,$$

Here, a_j is the average distance from the data point j to all other points in the same cluster, and b_j is the average distance from the data point j to all points in the nearest cluster. ASW is derived from mean of these values per group. Values near 1 indicate well-defined clusters with clear cohesion and separation, near 0 imply boundary objects possibly indicating overlap, and near -1 suggest potential misclassification with objects closer to neighboring clusters.

With $k = 9$ groups identified, Table 1 summarizes each cluster’s details: paper count (n_i), cohesion index (C_i), Average Silhouette Width (ASW_i), and the most representative paper, whose selection is based on the lowest cosine distance to the cluster centroid, helping to the identification of the main topic of the cluster. Some papers remain unallocated to a cluster due to a missing abstract, hence only the title has been retrieved, or due to word retention criteria based on entropy levels, omitting 69 observations lacking words at the specified level, amounting to 2.5% of the total.

Table 1. Most representative paper, numerosity, cohesion index and Average Silhouette Width for each identified cluster.

Cluster	Paper	n_i	C_i	ASW _{<i>i</i>}
1	Blavatsky, S., <i>Misinformation on the Russian-Ukrainian war: A case study</i> , 2023, ASU library	415	0.47	0.03
2	Hassen, B., T., El Bilali, H., <i>Impacts of the Russia-Ukraine War on Global Food Security: Towards More Sustainable and Resilient Food Systems?</i> , 2022, Foods	146	0.86	0.38
3	Poberezhets, V., <i>Healthcare crisis in Ukraine-worrying consequences of the Russian-Ukrainian war</i> , 2022, Croatian medical journal	361	0.61	0.12
4	Mbah, R. E., Wasum, D. F., <i>Russian-Ukraine 2022 War: A review of the economic impact of Russian-Ukraine crisis on the USA, UK, Canada, and Europe</i> , 2022, Advances in Social Sciences Research Journal	298	0.58	0.04
5	Zakeri, B., Paulavets, K., Barreto-Gomez, L., Echeverri, L.G., Pachauri, S., Boza-Kiss, B., Zimm, C., Rogelj, J., Creutzig, F., Ürge-Vorsatz, D. and Victor, D.G., <i>Pandemic, war, and global energy transitions</i> , 2022, Energies	121	0.83	0.30
6	Chowdhury, E. K., and Khan, I. I., <i>Reactions of global stock markets to the Russia-Ukraine war: An empirical evidence</i> , 2023, Asia-Pacific Financial Markets	203	0.73	0.19
7	Kyzym, M. O., Khaustova, V. Y., Shlykova, V. O., <i>War in Ukraine: analysis of the prerequisites, lessons of the infighting, and conclusions for the future</i> , 2022, Repository Simon Kuznets Kharkiv National University of Economics	776	0.37	-0.01
8	Fouéré, E., <i>Can the war in Ukraine revive the EU's enlargement agenda for the Western Balkans?</i> , 2022, CEPS	279	0.80	0.29
9	Luniachek, V., Ignatenko, L., <i>Transformation processes in secondary educational institutions at the time of the war (illustrated by the case of Kharkiv, Ukraine)</i> , 2022, Repository Simon Kuznets Kharkiv National University of Economics	77	0.82	0.35

Clusters 2, 5, and 9 are the smallest ones, with 146, 121, and 77 papers respectively. They exhibit an important homogeneity, reflected in their cohesion index values and ASW scores, indicating well-defined groups. Cluster 2 has the highest cohesion at 0.86, followed closely by cluster 5 at 0.83, while cluster 9 also shows strong cohesion at 0.82. Clusters 1 and 4 have low ASW values, pointing to a possible broader topic: this may also be due to the high number of papers belonging to them. Cluster 7 could be considered residual due to its low cohesion index value of 0.37 and negative ASW value. Figure 1 presents a wordcloud plot depicting the top 20 stemmed words in each identified cluster. The size of the wordclouds depends on the number of articles, while the size of individual words represents their frequency within the clusters. Additionally, multidimensional scaling determines clusters arrangement.

It is well known how fast and instant the spread of war-related news on social media is. This perspective is highlighted in cluster number 1, that could represent the ensemble of papers treating arguments related to online users and researches about the impact the war had on social media and the way they are perceived by the online community.

Food scarcity emerges as an important issue associated with warfare, as evidenced in cluster 2. The conflict profoundly impacts Ukraine's essential agricultural sector, notably the wheat cultivation. Another important aspect to consider is the healthcare one: it is widely acknowledged that hospitals have endured

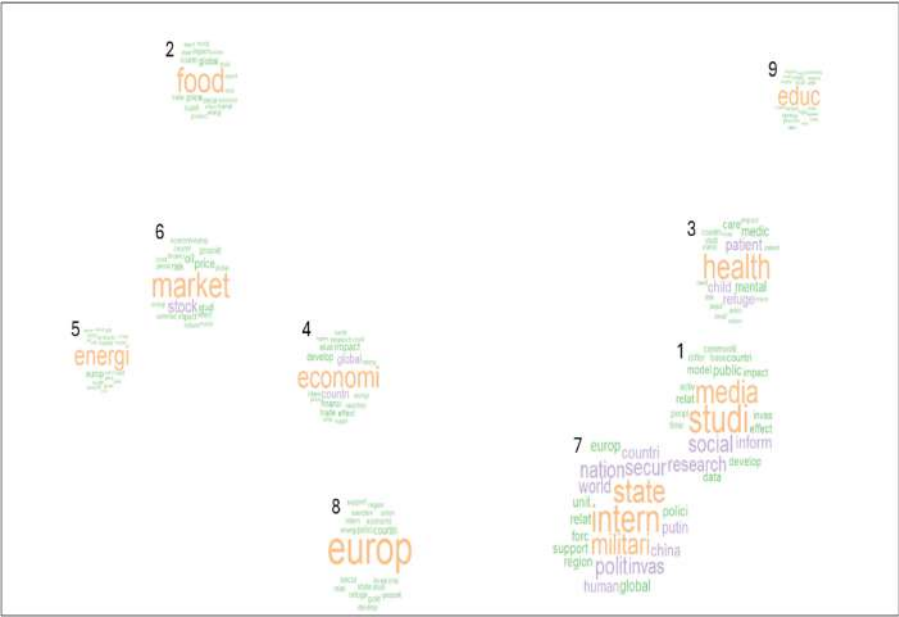


Fig. 1. Wordcloud plot of the topics identified in Ukrainian-Russian war related articles.

attacks since the onset of the conflict, worsening conditions for civilians, particularly children. This serves as the focal point identified within cluster 3.

Economic and financial consequences are grouped in clusters 4, 5, 6, each of them pertaining to a specific field. Cluster 4 contains papers about economic concerns that could be specific not only to the countries involved in the war, but in Europe and also in the entire globe. Group 5 portrays energy-related issues, primarily focusing on the ongoing impact of gas price repercussions that continue to affect the population. Cluster 6 encompasses the collective response of global and stock markets. This theme group closely parallels that of cluster 5, as well as the fourth one. It is noteworthy that the application of multidimensional scaling is fundamental in revealing the semantic proximity among these clusters.

The group number 7 has the highest frequency and can be characterized as the residual cluster, as it represents a topic that is not directly associated and surpasses the others in terms of prominence. It concerns military forces and possibly nationalism, with certain aspects typically associated with politics and security intertwined.

Cluster 8 foregrounds the exploration of geopolitical and diplomatic impacts resulting from crises, with a plethora of scholarly works delving into their far-reaching effects. Additionally, it appears that the aforementioned effects predominantly pertain to the economic realm. Indeed, this cluster shares semantic proximity with cluster 4.

The final group, designated as cluster 9, is about the exploration of educational aspects as a subject of research, emphasizing a notable concentration on student engagement and the prospective effects on the educational sphere.

The inclusion of hundreds of scientific outputs allowed to the identification of nine clusters, marking an increase of two compared to the prior study [1]. This divergence in outputs reflects the emergence of cluster number 1, formerly categorized within the residual category, and cluster number 5, associated with energy, previously encompassed within economic and financial themes.

4 Conclusion

This study presents a comprehensive analysis of a contemporary conflict, notably the Russia-Ukraine one. Through meticulous examination of scholarly research, we have gained deeper insights into the complex nature of these conflicts, spanning geopolitical tensions, economic repercussions, and societal impacts. By expanding on previous investigations and employing advanced clustering techniques, we have identified key thematic clusters, shedding light on critical issues such as the role of social media, economic challenges, and diplomatic implications. Language choice is crucial. We focused on English abstracts due to its dominance in science, but including languages like Ukrainian and Russian could enrich our insights. Expanding to multiple languages broadens our scope, capturing diverse scientific perspectives for a more inclusive research landscape understanding. The decision to analyze only abstracts was motivated by their efficiency in providing a quick overview of the topic treated in the article; we are aware that this approach may not fully capture the complexity of the full paper. By providing an overview of the scientific literature related to this contemporary conflict, we believe that this study will offer useful insights for policymakers, diplomats and stakeholders, helping them in informed decisions.

References

1. Scarso, L., Novelli, M., Violante, F.S.: Scientific abstracts analysis to uncover Ukrainian-Russian war insights: a text mining-based approach. In: Book of Short Papers ASA 2023 (2023)
2. Anderlucci, L., Montanari, A., Viroli, C.: The importance of being clustered: uncluttering the trends of statistics from 1970 to 2015. *Stat. Sci.* **34**(2), 280–300 (2019)
3. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **42**, 143–175 (2001)
4. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)



Competing Sets of Predictors in an Authorship Attribution Task: Most Frequent Words, Large Language Models and Correspondence Analysis

Andrea Sciandra^{1,2}(✉)  and Arjuna Tuzzi¹ 

¹ Department of Philosophy, Sociology, Pedagogy and Applied Psychology (FISPPA),
Università di Padova, Padua, Italy

arjuna.tuzzi@unipd.it

² Department of Statistical Sciences, Università di Padova, Padua, Italy
andrea.sciandra@unipd.it

Abstract. This study aims to assess the performance of different feature sets in an authorship attribution test using a large corpus of 76 contemporary Italian popular mystery novels by 16 authors. The feature sets include the dimensions derived from a large language model, the most frequent words, and the coordinates of correspondence analysis. Our analysis compares and contrasts the results obtained through these different vector representations in machine learning classification tasks. Although transformers have been shown to outperform other alternatives in previous works, in this case, correspondence analysis proves to be the winner of the challenge. The results support the hypothesis that specialized large corpora require tailor-made representations.

Keywords: authorship attribution · machine learning · correspondence analysis · transformers · popular Italian mystery novels

1 Introduction

The rise of large language models (LLMs) and the availability of transformers (Wolf et al., 2020) for vector representation of words and texts have significantly transformed the field of text mining. They have shown remarkable progress in various applications related to text classification tasks. However, the question of whether LLMs represent the ultimate solution to all problems remains. This is due to the opaque nature of this technology, which greatly limits the ability to produce explainable models. This study aims to evaluate the performance of LLMs in an authorship attribution task. The results will be compared and contrasted with those obtained through the most frequent words (MFWs), and a through correspondence analysis (CA, Greenacre, 1984; Murtagh, 2005) as proposed by Lebart (1997). This paper is an updated version of the paper presented at the conference Statistics for Data Science and Artificial Intelligence (Palermo - Italy, 11–12 April 2024; Sciandra & Tuzzi, 2024).

The corpus consists of 76 contemporary Italian popular mystery novels written by 16 different authors. Each author contributes a minimum of three and a maximum of six novels, making it a highly homogeneous collection in terms of text genre. The corpus was carefully assembled with the support of the publishing house, which actively

participated in selecting the most significant authors and works. A tokenization method that only considers alphabetic sequences (with all other characters treated as separators, and numbers removed) and basic normalization that converts all words to lowercase was employed. The corpus contains just under five million words (4,845,966 word tokens) and has a vocabulary of slightly over 100 thousand distinct words (109,643 word types). The novels vary in length from 22,402 to 123,457 words (see Fig. 1), resulting in considerable variability in both the length of the works and the contribution of each author to the corpus. The corpus has a type-token ratio of 2.2% (indicating an average word type frequency of 44.2) and a percentage of *hapax legomena* at 38.1% of the vocabulary, demonstrating a desirable high level of redundancy.

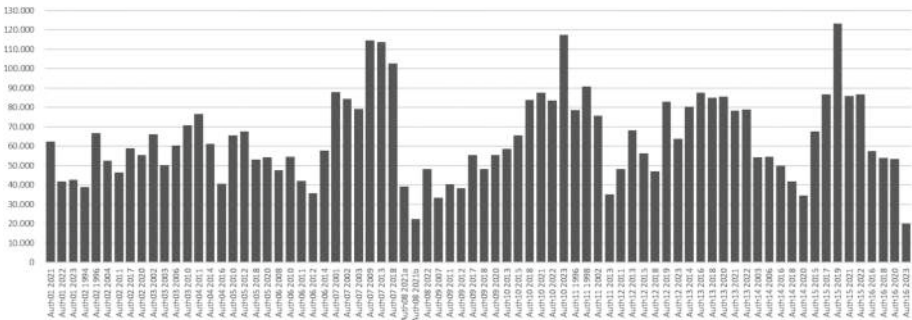


Fig. 1. Dimensions in word tokens of novels.

Since it exhibits homogeneity in terms of text genre and comprises novels of undisputed authorship, our corpus serves as a solid foundation for evaluating the efficacy of a classification pipeline in an authorship attribution task. The aim of this study is to assess the performance of various machine learning-based classification methods, each using different approaches to feature representation. We compared vectorization methods using modern transformers, those based on the MFWs, and those provided by traditional CA. The texts were divided into 9540 chunks of equal size (500 words each) as we utilized an LLM model capable of processing texts with a maximum length of 512 tokens. Performance was assessed based on the accuracy of attributing each chunk in the test set to the correct author.

2 Methods

In terms of ML-based classification methods, we selected three models: Random Forests (RF; Breiman, 2001), Extreme Gradient Boosting (XGBoost; Chen et al., 2019), and Support Vector Machine with radial basis function kernel (SVM-RK; Scholkopf et al., 1997). Our idea is to provide informative validations using bagging and boosting techniques, as well as a popular classifier suitable to capture non-linear relationships (e.g., overlapped classes). The training set was created by randomly selecting 80% of the chunks from each author, while the remaining chunks were used for the test set. Each model underwent a 5-fold cross-validation, repeated 30 times. The results were assessed

by analyzing the accuracy, the unweighted Kappa statistic, and a one-sided test to determine if the accuracy exceeded the ‘no information rate’ (the largest class percentage in the data: 12%; Kuhn, 2008). The aim is to highlight respectively: the proportion of correctly classified chunks, the level of agreement between predicted and actual authors (in other words, the performance of each classifier compared to the baseline of a random classifier), and the overall significance of each model.

We selected word types with number of occurrences ≥ 76 (one occurrence per novel on average) and started from the (5068 x 9540) Term-Document Matrix (TDM) with (relative) frequencies as weights. After calculating the chi-square distance and Singular Value Decomposition (SVD), each text (word) becomes a column (row) vector in the 5067-dimensional space of CA orthogonal axes.

Subsequently, we used transformers to represent texts, specifically utilizing the BERT Base Italian Uncased LLM pre-trained vectors (available at <https://huggingface.co/dbmdz/bert-base-italian-uncased>). BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019) provides contextualized embeddings, as they consider the context in which a word is used. As a result, they dynamically generate word representations informed by the surrounding words. In this study, we concatenated the 12 BERT layers representing the same token and used the mean to aggregate the embeddings from different tokens to represent a text.

As the BERT-base mapping involves 768 vectors, we chose three different settings:

- to use all BERT vectors, compared to an equal number of CA dimensions, and an equal number of words and n-grams (in this setting: the 200 MFWs - basically function words - together with the 500 most frequent character n-grams, and the 68 most frequent word n-grams).
- To use a small number of features, equal to the number of novels (76). Therefore we selected the 76 MFWs, the first 76 axes of CA (the explained variance for the first 76 axes is about 13%). Then, to reduce BERT vectors dimensionality, we applied a principal component analysis (PCA), extracting the first 76 components (these components account for approximately 82% of the cumulative proportion of variance of embeddings obtained via LLM).
- For the purpose of testing an intermediate number of features, we decided to select the 500 MFWs, the first 500 axes of the CA and, similarly to the above, to use PCA to extract the first 500 principal components from BERT’s vectors (99% of the cumulative proportion of variance).

3 Discussion and Conclusions

Table 1 presents the results of the models applied to the test set.

It is noteworthy that models using CA consistently outperformed those using a LLM. The XGBoost model, which includes 76 features derived from CA, emerges as the top performer with an accuracy of 99.33%. This result is particularly remarkable given the relatively limited number of features used. The Kappa statistic provides further validation, confirming the XGBoost model with 76 CA coordinates as the most accurate. All models demonstrate statistical significance based on the one-sided test, exceeding the accuracy threshold of the no information rate (NIR). These results highlight the superior classification performance of CA features within this specific corpus.

The efficacy of CA in distinguishing between novel authors becomes evident when projecting the chunks belonging to the test set onto a Cartesian plane. In this instance, dimensions 7–8 and 11–12 are used. In the first plane, author number 1 is isolated by dimension 7 and author 15 is isolated by dimension 8. In the second plane, author number 14 is isolated by dimension 11 (Fig. 2).

While previous studies (Iyer & Vosoughi, 2020; Jones et al., 2022; Ai et al., 2022; Rodella et al., 2024) have demonstrated the superiority of transformers over alternative methods, in this case CA emerges as the better option. However, in other multilingual investigations, the adoption of BERT failed to enhance authors’ classification performance when compared to MFWs (Škorić et al., 2022). Our hypothesis posits that the effectiveness of transformers depends on the specific application context. Transformers, which were developed by training LLMs on diverse and extensive databases, provide a robust representation for general text classification. However, their performance may

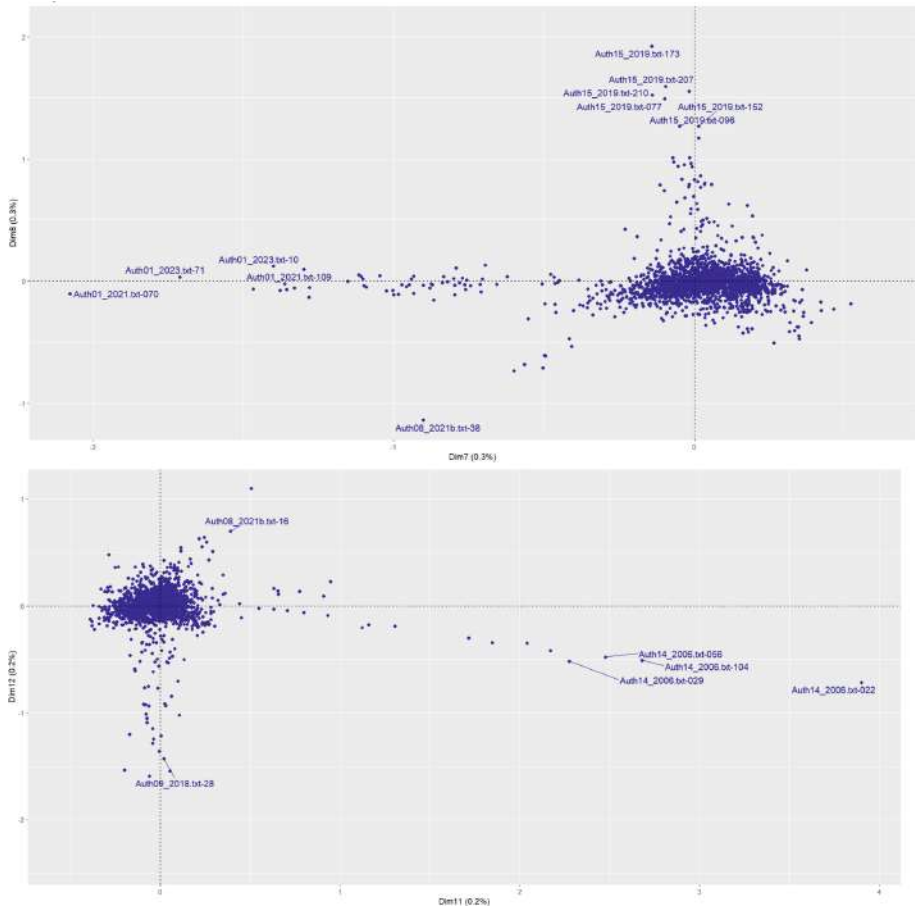


Fig. 2. Projection of the chunks belonging to the test set onto the Cartesian planes formed by dimensions 7–8 (above) and dimensions 11–12 (below) of the CA.

decrease when dealing with non-English languages, large and complex texts, or specialized language and genre corpora. In such cases, methods derived from the analysis of the corpus's textual data itself may be more effective, as they use information from the training set. Still, it is necessary to rigorously test our hypothesis across different corpora and using alternative machine learning methods.

It is worth noting that some LLMs do not disclose the materials on which they were trained, which presents a challenge for research in this domain. Despite the abundance of open-source tools, resources, and code available online, there is still a lack of comprehensive discussion, documentation, and scientific publications. Therefore, navigating through the multitude of options without clear criteria to evaluate their suitability for specific purposes and contexts of use can be a significant challenge.

Table 1. Comparison of author classification performance: LLM, words and CA (Diagnostics: Accuracy, unweighted Kappa statistic, and a one-sided test p -value to verify if the accuracy is greater than NIR). The number of features used appears in parentheses next to the model name. In bold are highlighted the models that achieve the best accuracy result for each feature set.

ML Model	Accuracy	Kappa	P-Value (Acc > NIR)
LLM – RF (768)	0.7629	0.7416	<2.2e-16
LLM – XGBoost (768)	0.8933	0.8845	<2.2e-16
LLM – SVM-RK (768)	0.9309	0.9253	<2.2e-16
LLM – RF (76)	0.7438	0.7205	<2.2e-16
LLM – XGBoost (76)	0.8521	0.8397	<2.2e-16
LLM – SVM-RK (76)	0.9052	0.8974	<2.2e-16
LLM – RF (500)	0.6892	0.6603	<2.2e-16
LLM – XGBoost (500)	0.8799	0.8698	<2.2e-16
LLM – SVM-RK (500)	0.9593	0.9559	<2.2e-16
Words – RF (768)	0.7412	0.7174	<2.2e-16
Words – XGBoost (768)	0.9005	0.8923	<2.2e-16
Words – SVM-RK (768)	0.9454	0.9409	<2.2e-16
Words – RF (76)	0.5448	0.4995	<2.2e-16
Words – XGBoost (76)	0.6330	0.6015	<2.2e-16
Words – SVM-RK (76)	0.6696	0.6416	<2.2e-16
Words – RF (500)	0.8505	0.8377	<2.2e-16
Words – XGBoost (500)	0.9387	0.9337	<2.2e-16
Words – SVM-RK (500)	0.9526	0.9487	<2.2e-16

(continued)

Table 1. (continued)

ML Model	Accuracy	Kappa	P-Value (Acc > NIR)
CA – RF (768)	0.9866	0.9855	<2.2e-16
CA – XGBoost (768)	0.9923	0.9916	<2.2e-16
CA – SVM-RK (768)	0.9866	0.9855	<2.2e-16
CA – RF (76)	0.9856	0.9844	<2.2e-16
CA – XGBoost (76)	0.9933	0.9928	<2.2e-16
CA – SVM-RK (76)	0.9923	0.9916	<2.2e-16
CA – RF (500)	0.9871	0.9861	<2.2e-16
CA – XGBoost (500)	0.9907	0.9900	<2.2e-16
CA – SVM-RK (500)	0.9902	0.9894	<2.2e-16

References

Ai, B., Wang, Y., Tan, Y., Tan, S.: Whodunit? Learning to contrast for authorship attribution. *arXiv. arXiv:2209.11887* (2022)

Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: *xgboost: Extreme gradient boosting*. R package version 0.81.0.1, pp. 1–4 (2019)

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/N19-1423>

Greenacre, M.J.: *Theory and Applications of Correspondence Analysis*. Academic Press, London (1984)

Iyer, A., Vosoughi, S.: Style Change Detection Using BERT. *CLEF (Working Notes)* **93**, 106 (2020)

Jones, K., Nurse, J.R.C., Li, S.: Are you robert or RoBERTa? Deceiving online authorship attribution models using neural text generators. *Proc. Int. AAAI Conf. Web Soc. Media* **16**, 429–440 (2022). <https://doi.org/10.1609/icwsm.v16i1.19304>

Kuhn, M.: Building predictive models in R using the caret package. *J. Stat. Softw.* (2008). <https://doi.org/10.18637/jss.v028.i05>

Lebart, L.: Correspondence analysis, discrimination, and neural networks. In: Hayashi, C., Yajima, K., Bock, H.H., Ohsumi, N., Tanaka, Y., Baba, Y. (eds.) *Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 423–430. Springer, Tokyo (1998). https://doi.org/10.1007/978-4-431-65950-1_47

Murtagh, F.: *Correspondence analysis and data coding with Java and R*. CRC Press (2005)

Rodella, I., Sciandra, A., Tuzzi, A.: Analysis of Marie Skłodowska-Curie Actions (MSCA) evaluations and models for predicting the success of proposals. In: Dister, A., Longrée, D. (eds.) *JADT 2024 Mots comptés, textes déchiffrés*, pp. 783–792 (2024)

Scholkopf, B., et al.: Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **45**(11), 2758–2765 (1997)

Sciandra, A., Tuzzi, A.: Can correspondence analysis challenge transformers in authorship attribution tasks? In: Plaia, A., Egidi, L., Abbruzzo, A. (eds.) *Proceedings of the Statistics and Data*

Science 2024 Conference – New perspectives on Statistics and Data Science, pp. 362–367 (2024)

Škorić, M., Stanković, R., Ikonić Nešić, M., Byszuk, J., Eder, M.: Parallel stylometric document embeddings with deep learning based language models in literary authorship attribution. *Mathematics* **10**(5), 838 (2022)

Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>



Clustering Models for Time Series Analysis in Confocal Microscopy

Ivan Sciascia^(✉), Andrea Crosino, and Andrea Genre

Department of Life Sciences and Systems Biology, Università degli Studi di Torino,
viale Mattioli 25, 10125 Turin, Italy
`ivan.sciascia@unito.it`

Abstract. We analyze the time series of two experimental sets to verify whether there are periodicities in the emission of calcium spiking. We use signal classification methods with clustering techniques.

Keywords: calcium spiking · spectral analysis · periodicity · mycorrhiza

1 Calcium Spiking Time Series

We analyzed series of calcium peak emission with the confocal microscope which can be used for the live imaging of cellular and physiological processes as they develop in living cells. The calcium-mediated signaling is a common mechanism of signal transduction across eukaryotes which is studied in root symbiotic interactions with nitrogen-fixing bacteria and arbuscular mycorrhizal fungi [3]. The perception of microbe-derived molecules by root epidermal cells triggers oscillations in nuclear calcium concentration, known as calcium spiking, which with systemic cellular responses allow the establishment of symbiosis [1, 2]. In confocal microscopy the analysis of two fluorophores that label protein constructs gives rise to the FRET (Förster Resonance Energy Transfer) phenomenon: resonance energy transfer between fluorophores, that occurs when donor fluorophore emission wavelength and acceptor fluorophore absorption wavelength substantially overlap. In this study we analyzed the calcium peaks recorded in root epidermal cells after their inclusion with short-chain chito-oligosaccharides, the so-called Myc factors, and with NOD factors. Such calcium spike traces are characterized by a series of calcium concentration spikes that begin a few minutes after the application of Myc and NOD factors and continue for over 10 min [4]. The peak sequence is markedly irregular for the Myc factor, compared to the more regulated peaks described for the NOD factors. In this work we compare the responses of the two factors trying to find a periodicity in the calcium peaks with cluster analysis classification methods. The compounds able to stimulate symbiosis between legumes and nitrogen-fixing bacteria are commonly known as NOD-factors. Such Nod-factors are mainly composed of lipochitoooligosaccharides which are chitin-derived molecules, while chitoooligosaccharides (COs) are fungi-released molecules, generally associated with arbuscular

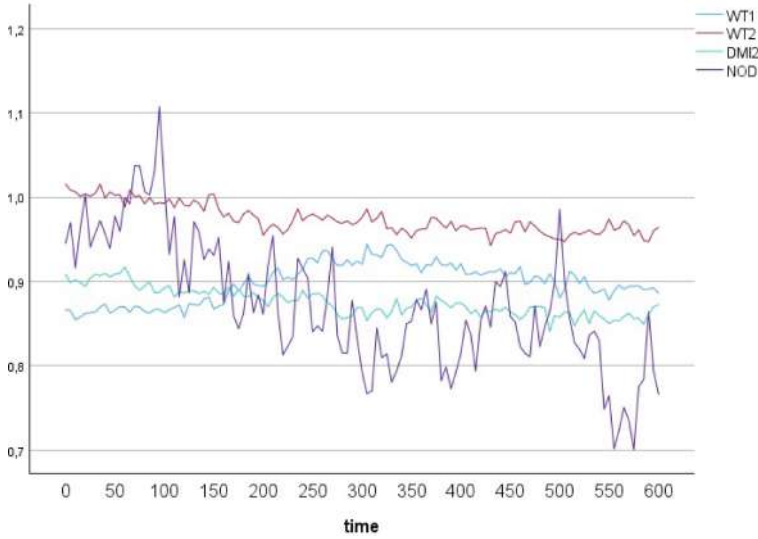


Fig. 1. Time series graph

mycorrhizal symbiosis. The subsequent analyses are run both on wild type (WT) *Medicago truncatula* roots and on mutant roots unable to perceive COs (DMI2) after CO treatment.

2 Cluster Analysis

In this section we analyze the peaks for myc factors expressed by three experimental sets, WT1, WT2 and DMI2 and the peaks for NOD factors. The overlapping series are described in Fig. 1 where a similarity between the myc factor series is noted while the NOD factor induces a series with a greater amplitude than the FRET intensity ratio axis. The analysis of the periodogram is described in Fig. 2 and shows how it is possible to understand a difference in the activation period of a possible periodicity at different times for the WT1 trace and the NOD trace that we compare. The correlation analysis is expressed in Fig. 3 and shows how we can identify the best correlation between the DMI2 trace and the WT2 trace (Pearson coefficient 0.797, $\alpha = 0.01$) and between the DMI2 trace and the NOD (Pearson coefficient 0.700, $\alpha = 0.01$).

The classification analysis is carried out to explore whether there are comparably frequent time intervals or calcium peaks in the time series. In this way we could have an indication of the periodicity of the series of peaks. The k-means classification results are described in Table 1.

The mode cluster center in Table 1 represents for every dataset the cluster with great number of cases and relative cluster mean assuming the division of the time series into four clusters. The results of the time series analysis,

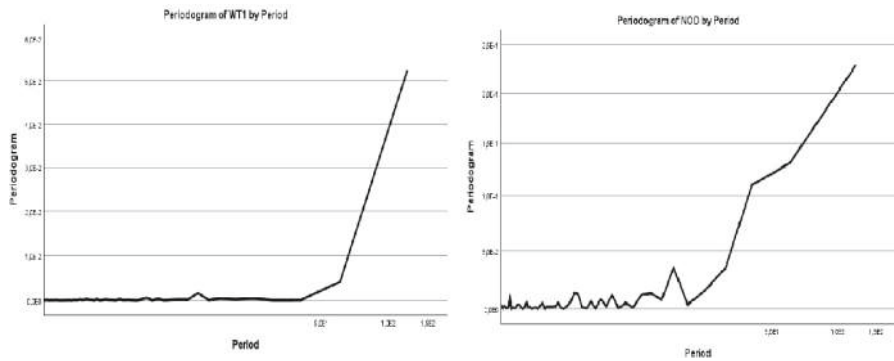


Fig. 2. Compared periodograms

		Correlations			
		WT1	WT2	DMI2	NOD
WT1	Pearson Correlation	1	-,609**	-,586**	-,590**
	Sig. (2-tailed)		<,001	<,001	<,001
	N	121	121	121	121
WT2	Pearson Correlation	-,609**	1	,797**	,648**
	Sig. (2-tailed)	<,001		<,001	<,001
	N	121	121	121	121
DMI2	Pearson Correlation	-,586**	,797**	1	,700**
	Sig. (2-tailed)	<,001	<,001		<,001
	N	121	121	121	121
NOD	Pearson Correlation	-,590**	,648**	,700**	1
	Sig. (2-tailed)	<,001	<,001	<,001	
	N	121	121	121	121

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 3. Time series correlations

Table 1. Classification for WT1, WT2, DMI2, NOD

Series	Mode cluster center	Number of cases
WT1	0.913	40
WT2	0.958	48
DMI2	0.885	42
NOD	0.85	53

correlations and k-means cluster classification demonstrate that there are clues to the similarity of the periodicity of peaks induced by NOD and Myc factors which could be better explored with the complementary use of techniques such as cluster analysis and spectral analysis of signals.

References

1. Bonfante, P., Requena, N.: Dating in the dark: how roots respond to fungal signals to establish arbuscular mycorrhizal symbiosis. *Curr. Opin. Plant Biol.* **14**, 451–457 (2011)
2. Genre, A., et al.: Short chain chitin oligomers from arbuscular mycorrhizal fungi trigger nuclear Ca^{2+} spiking in *Medicago truncatula* roots and their production is enhanced by strigolactone. *New Phytol.* **198**, 190–202 (2013)
3. Maillet, F., et al.: Fungal lipochitooligosaccharide symbiotic signals in arbuscular mycorrhiza. *Nature* **469**, 58–64 (2011)
4. Russo, G., Spinella, S., Sciacca, E., Bonfante, P., Genre, A.: Automated analysis of calcium spiking profiles with CaSA software: two case studies from root-microbe symbioses. *BMC Plant Biol.* **13**, 224 (2013)



Minimax Rates for Wasserstein Deconvolution of Regular Distributions with Known Ordinary Smooth Errors

Catia Scricciolo^(✉)

Università degli Studi di Verona, Via Cantarane 24, 37129 Verona, Italy
catia.scricciolo@univr.it

Abstract. We study the problem of univariate distribution function deconvolution under Wasserstein metrics, with known and ordinary smooth error distributions. We fill a gap present in the literature extending the result on the upper bound rates under Wasserstein metrics for a recently proposed isotone distribution function estimator to the case when the mixing distribution has some level of regularity either in a Hölder or in a Sobolev scale. The found rates turn out to be minimax-optimal (up to a logarithmic factor) over the full scale of values of the smoothness index $\beta > 0$ of the Fourier transform of the error distribution under the 1-Wasserstein distance, while, for Wasserstein metrics of order $p > 1$, these rates are known to be minimax-optimal only for $\beta \leq \frac{1}{2}$.

Keywords: deconvolution · minimax rates · ordinary smooth errors · Wasserstein metrics

1 Introduction

We study the problem of univariate distribution deconvolution under Wasserstein metrics, with known and ordinary smooth error distributions. Suppose we observe real-valued random variables (r.v.'s) $Y_1, \dots, Y_n$ such that Y_i is a signal X_i additively contaminated with an error ε_i , in symbols

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where the sequences $(X_i)_{i \in \mathbb{N}}$ and $(\varepsilon_i)_{i \in \mathbb{N}}$ are independent, the X_i 's are independent and identically distributed (i.i.d.) according to an unknown distribution μ_X , while the ε_i 's are i.i.d. according to a known distribution μ_ε . We assume that μ_ε is ordinary smooth, in the sense that, given a real number $\beta > 0$, there exist constants $c_1, c_2 > 0$ such that the Fourier transform $\hat{\mu}_\varepsilon(\cdot) = \int_{\mathbb{R}} e^{iu} \mu_\varepsilon(du)$ of the error distribution μ_ε verifies the condition (Table 1).

$$c_1 |t|^{-\beta} \leq |\hat{\mu}_\varepsilon(t)| \leq c_2 |t|^{-\beta}, \quad \text{for large } |t|.$$

Table 1. Notable ordinary smooth error distributions

Distribution	$\hat{\mu}_\varepsilon$	Smoothness index β
Laplace	$(1 + t^2)^{-1}$	2
Symmetric Linnik	$(1 + t ^\beta)^{-1}$	$0 < \beta \leq 2$
Symmetrized Gamma	$(1 + t^2)^{-\frac{\beta}{2}}$	$\beta > 0$
Symmetrized χ_k^2	$(1 + 4t^2)^{-\frac{k/2}{2}}$	$k/2, \quad k \in \mathbb{N}$

The interest is in recovering the mixing distribution μ_X on the basis of a random sample $Y_1, \dots, Y_n$ attaining minimax-optimal convergence rates under Wasserstein metrics. Let $\mathcal{P}(\mathbb{R})$ be the set of all probability measures on $\mathbb{R}$. For $1 \leq p < \infty$, let $\mathcal{P}_p(\mathbb{R})$ be the set of probability measures μ on $\mathbb{R}$ having finite p th absolute moments $M_p(\mu) := \int_{\mathbb{R}} |x|^p \mu(dx) < \infty$, in symbols, $\mathcal{P}_p(\mathbb{R}) := \{\mu \in \mathcal{P}(\mathbb{R}) : M_p(\mu) < \infty\}$. The Wasserstein distance of order p between two probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ is defined as

$$W_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{\mathbb{R} \times \mathbb{R}} |x - y|^p \pi(dx, dy) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ stands for the set of joint distributions (*couplings*) on $\mathbb{R} \times \mathbb{R}$ with marginals μ and ν .

The problem has been investigated by, among others, Caillerie *et al.* [1], Dedecker *et al.* [2], Dedecker and Michel [3]. In particular, Dedecker *et al.* [2] propose an estimator $\hat{F}_n$ for the distribution function F_X associated with the unknown mixing distribution μ_X that is a genuine cumulative distribution function. They begin by considering a preliminary estimator $\tilde{F}_n$, which is not necessarily a non-decreasing function, defined as the integral of the classical deconvolution kernel density estimator, see, *e.g.*, Ch. 10 by Delaigle in Yi *et al.* [4],

$$\hat{F}_n(x) := \frac{1}{nh} \int_{-\infty}^x \sum_{i=1}^n \tilde{k}_h \left(\frac{u - Y_i}{h} \right) du, \quad x \in \mathbb{R},$$

where

$$\tilde{k}_h(x) := \frac{1}{2\pi} \int_{\mathbb{R}} e^{itx} \frac{\hat{k}(t)}{\hat{\mu}_\varepsilon(-t/h)} dt, \quad x \in \mathbb{R}.$$

For a kernel k such that $\hat{k}$ is symmetric, the function $\tilde{k}_h(\cdot)$ coincides with the classical deconvolution kernel because

$$\tilde{k}_h(x) := \frac{1}{2\pi} \int_{\mathbb{R}} e^{itx} \frac{\hat{k}(t)}{\hat{\mu}_\varepsilon(-t/h)} dt = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-isx} \frac{\hat{k}(s)}{\hat{\mu}_\varepsilon(s/h)} ds, \quad x \in \mathbb{R}.$$

Then, given $\rho \geq 0$, the estimator $\hat{F}_n$ is defined as the approximate (up to an error ρ) minimizer over all distribution functions (d.f.'s) F of the quantity

$\int_{\mathbb{R}} |x|^{p-1} |(F - \hat{F}_n)(x)| \, dx$, that is,

$$\int_{\mathbb{R}} |x|^{p-1} |(\tilde{F}_n - \hat{F}_n)(x)| \, dx \leq \int_{\mathbb{R}} |x|^{p-1} |(F - \hat{F}_n)(x)| \, dx + \rho \quad \text{for all d.f.'s } F. \quad (1)$$

If $\rho = 0$, then $\tilde{F}_n$ is the best approximation. In what follows, the choice $\rho \equiv \rho_n \leq n^{-1/2}$ would suffice for the purpose. Therefore, the estimator $\tilde{\mu}_n$ of μ_X is defined as the probability measure associated with the distribution function $\tilde{F}_n$.

Under a tail condition on the distribution of Y and some regularity conditions on $\hat{\mu}_\varepsilon$, Dedecker *et al.* [2] have obtained an upper bound of the order $O(n^{-1/[2p+(2\beta-1)+]})$, up to a logarithmic factor, on the rate of convergence of $\tilde{\mu}_n$ under the p th Wasserstein risk, for any $p \geq 1$, cf. *ibid.*, Theorem 3.1, $\mathbb{E}_Y^n[W_p^p(\tilde{\mu}_n, \mu_X)] \leq Cn^{-p/[2p+(2\beta-1)+]}$. In particular, for $p = 1$, this upper bound matches with the lower bound $n^{-1/[(2\beta \vee 1)+1]}$ obtained by the same authors in the form $n^{-1/(2\beta+1)}$ and refined by Rousseau and Scricciolo [6], see Theorem 5.1. Since no assumptions on μ_X , except for moment conditions, are postulated, an open question remains the extension of the previous result to the case when the mixing distribution is regular. In this note, our goal is to derive minimax-optimal rates of convergence, at least for $p = 1$, that are attained by the estimator $\tilde{\mu}_n$ when the unknown distribution μ_X is assumed to be regular.

The note is organized as follows: in Sect. 2 we expose the main result, while in Sect. 3 we report some final remarks and comments.

2 Main Result

In this section, we present the result on rates of convergence under p -Wasserstein metrics for the estimator $\tilde{\mu}_n$ to recover the unknown mixing distribution μ_X assumed to be Sobolev regular in the sense made precise hereafter.

Assumption 1. For $p \geq 1$, the distribution $\mu_X \in \mathcal{P}_p(\mathbb{R})$ is such that there exists $\alpha > 0$ for which

$$\int_{\mathbb{R}} |t|^\alpha |\hat{\mu}_X(t)| \, dt < \infty \quad \text{and} \quad (|\cdot|^{\alpha+r} \hat{\mu}_X^{(r)}) \in L^2(\mathbb{R}) \quad \text{for } r = 0, \dots, [p],$$

where $[p]$ denotes the least integer greater than or equal to p and $\hat{\mu}_X^{(r)}$ stands for the r th derivative of the Fourier transform $\hat{\mu}_X$ of μ_X .

To state the main result, we introduce some more notation. Let

$$r_\varepsilon(t) := \frac{1}{\hat{\mu}_\varepsilon(t)} \quad \text{for all } t \in \mathbb{R} \text{ such that } \hat{\mu}_\varepsilon(t) \neq 0$$

and, for $\ell \in \{0\} \cup \mathbb{N}$, let $r_\varepsilon^{(\ell)}$ denote the ℓ th derivative of r_ε . For $x \geq 0$, let $H_Y(x) := P(|Y| \geq x)$ be the tail probability of the distribution of $|Y|$.

The following result extends Theorem 3.1 of Dedecker *et al.* [2] to the case where the mixing distribution μ_X is regular and has an α -Sobolev regular density, thus providing an upper bound on the rate of convergence for the estimator $\tilde{\mu}_n$ under a p -Wasserstein metric of any order $p \geq 1$.

Theorem 1. Let $\tilde{\mu}_n$ be the isotone estimator corresponding to $\rho \equiv \rho_n \leq n^{-1/2}$. For $p \geq 1$, let μ_X satisfy Assumption 1. Denoted by m_0 the least integer strictly greater than $p + 1/2$, assume that r_ε is m_0 -times continuously differentiable,

$$S_Y := \int_0^\infty x^{p-1} \sqrt{H_Y(x)} dx < \infty, \quad \sup_{|t| \leq 2} |r_\varepsilon^{(m_0)}(t)| < \infty, \quad (2)$$

and there exist constants $\beta, c > 0$ such that

$$\forall \ell \in \{0, \dots, m_0 - 1\}, \quad |r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^\beta, \quad t \in \mathbb{R}.$$

Then, there exists a constant $C > 0$ such that

$$\mathbb{E}_Y^n[W_p^p(\tilde{\mu}_n, \mu_X)] \leq C\psi_n, \quad (3)$$

where $\psi_n \equiv \psi_n(\alpha, \beta, p) := n^{-\frac{\alpha+p}{2(\alpha+p)+(2\beta-1)+}} (\log n)^{\mathbf{1}_{\{1/2\}}(\beta)/2}$.

Proof. The proof follows the same track as the proof of Theorem 3.1 in Dedecker *et al.* [2]. In what follows, we take $k \in L^1(\mathbb{R})$ to be a symmetric, non-negative kernel with $\int_{\mathbb{R}} |x|^p k(x) dx < \infty$, whose Fourier transform is supported on $[-1, 1]$ and is $[p]$ -times differentiable with Lipschitz $[p]$ th derivative. For a kernel bandwidth $0 < h < 1$, let $k_h(\cdot) := h^{-1}k(\cdot/h)$ be the re-scaled version. The starting point is the following upper bound on $W_p^p(\tilde{\mu}_n, \mu_X)$ which is the sum of the bias of $\tilde{\mu}_n$ and a random term:

$$W_p^p(\tilde{\mu}_n, \mu_X) \leq 2^{p-1} [W_p^p(\mu_X * k_h, \mu_X) + W_p^p(\tilde{\mu}_n, \mu_X * k_h)].$$

The bias is controlled using a modification of Lemma C.2 of Rousseau and Scricciolo [5], which is not reported here for space constraints, so that

$$W_p^p(\mu_X * k_h, \mu_X) \leq \int_{\mathbb{R}} |x|^{p-1} |(F_X * k_h - F_X)(x)| dx = O(h^{\alpha+p}).$$

The random term can be controlled as in Dedecker *et al.* [2], using Proposition 3.1. We warn the reader that in what follows the constant C may change from line to line. We have

$$\mathbb{E}_Y^n[W_p^p(\tilde{\mu}_n, \mu_X * k_h)] \leq \frac{C}{\sqrt{n}} \sum_{j=1}^4 A_j,$$

where, defined for $h < 1$ the integral

$$I_\ell := \int_{|x| \leq 1/h} \frac{|r_\varepsilon^{(\ell)}(x)|^2}{x^2} \mathbf{1}_{[-1, 1]^c}(x) dx, \quad \ell = 0, \dots, m_0 - 1,$$

we have

$$\begin{aligned} A_1 &:= S_Y \times \sup_{|t| \leq 2} \sum_{\ell=0}^1 |r_\varepsilon^{(\ell)}(t)| = O(1), & A_2 &:= \sup_{|t| \leq 2} \sum_{\ell=0}^{m_0} |r_\varepsilon^{(\ell)}(t)| = O(1), \\ A_3 &:= \sqrt{M_{2p-1/2}(\mu_Y) I_0} = O \left(h^{-(\beta-1/2)_+} \left(\log \frac{1}{h} \right)^{\mathbf{1}_{\{1/2\}}(\beta)/2} \right), \\ A_4 &:= \sqrt{\sum_{\ell=0}^{m_0-1} I_\ell} = O \left(h^{-(\beta-1/2)_+} \left(\log \frac{1}{h} \right)^{\mathbf{1}_{\{1/2\}}(\beta)/2} \right). \end{aligned}$$

Putting all terms together, we get that

$$\mathbb{E}_Y^n[W_p^p(\tilde{\mu}_n, \mu_X)] \leq \frac{1}{\sqrt{n}} + 2^{p-1} h^{\alpha+p} + \frac{C}{\sqrt{n}} h^{-(\beta-1/2)_+} \left(\log \frac{1}{h} \right)^{\mathbf{1}_{\{1/2\}}(\beta)/2},$$

which, taking $h \equiv h_n = n^{-\frac{1}{2(\alpha+p)+(2\beta-1)_+}}$, gives the rate in (3). $\square$

Remark 1. For $p = 1$, the result holds true also if the mixing density is locally Hölder, in the sense that $\mu_X \in \mathcal{P}_1(\mathbb{R})$ has density f_X such that there exist $\alpha > 0$ and $L \in L^1(\mathbb{R})$ so that the derivative $f_X^{(\ell)}$ of order $\ell := \lfloor \alpha \rfloor$ (the greatest integer strictly smaller than α) exists and is locally Hölder of order $(\alpha - \ell)$, in symbols,

$$|f_X^{(\ell)}(x + \delta) - f_X^{(\ell)}(x)| \leq L(x) |\delta|^{\alpha-\ell} \text{ for every } \delta, x \in \mathbb{R}.$$

Under this assumption, in fact, using Lemma C.1 of Rousseau and Scricciolo [5], it can be shown that the bias $W_1(\mu_X * k_h, \mu_X) = O(h^{\alpha+1})$.

Remark 2. For $p = 1$, the upper bound

$$\psi_n(\alpha, \beta, 1) = n^{-\frac{\alpha+1}{2(\alpha+1)+(2\beta-1)_+}} (\log n)^{\mathbf{1}_{\{1/2\}}(\beta)/2}$$

matches (up to a logarithmic factor when $\beta = 1/2$) with the lower bound found by Rousseau and Scricciolo (2023), Theorem 5.1, for dimension $d = 1$. It is interesting to note that in this case (*i.e.*, $p = 1$) the regularity of the mixing distribution is $\alpha + 1$, which is the regularity of the mixing density augmented by one unit. This corresponds to estimating the primitive of the density which is, in fact, $(\alpha + 1)$ -times differentiable. When $0 < \beta \leq 1/2$, the rate $n^{-1/2}(\log n)^{\mathbf{1}_{\{1/2\}}(\beta)/2}$ is the same for all $p \geq 1$, thus the open question remains whether the upper bound in (3) is minimax-optimal also when $\beta > 1/2$ for $p > 1$.

Example 1 (Laplace error distribution). When the error distribution has a standard Laplace density $f_\varepsilon(u) = e^{-|u|}/2$, $u \in \mathbb{R}$, if the mixing density is either α -Sobolev regular or locally Hölder in the sense made precise in Remark 1, then the minimax-optimal rate of convergence under the 1-Wasserstein distance is $n^{-(\alpha+1)/(2\alpha+5)}$. This rate is attained by the estimator $\tilde{\mu}_n$ corresponding to an error $\rho = O(n^{-1/2})$. If the mixing distribution is assumed to satisfy only moment conditions and is not necessarily Lebesgue absolutely continuous (a case that can be regarded as corresponding to $\alpha = 0$), then the minimax-optimal rate reduces to $n^{-1/5}$, as pointed out by Dedecker *et al.* [2]. $\square$

The upper bound given in Theorem 1 reduces to the upper bound of Theorem 3.1 in Dedecker *et al.* [2] when the mixing distribution is assumed to satisfy only moment conditions and is not necessarily Lebesgue absolutely continuous (a case that can be thought to correspond to $\alpha = 0$). In the latter case, for $p > 1$ and $\beta > 1/2$, there is, in fact, a mismatch between the lower bound $n^{-1/(2\beta+1)}$ and the upper bound $n^{-1/(2p+2\beta-1)}$ given by Dedecker *et al.* [2] in Theorem 4.1, thus leaving open the question of minimax-optimal rates.

3 Final Remarks

In this note, we have considered the problem of univariate distribution deconvolution under Wasserstein metrics of any order $p \geq 1$, when the error distribution is completely known and ordinary smooth, while the mixing distribution is assumed to be regular, in particular Sobolev smooth. Given a sample of n observations from the convolved distribution, we have derived an upper bound on the rate of convergence for the estimator $\tilde{\mu}_n$ recently proposed by Dedecker *et al.* [2], which is of the order $O(n^{-[(\alpha/p)+1]/[2(\alpha+p)+(2\beta-1)+]})$, up to a logarithmic factor. This rate is minimax-optimal under the 1-Wasserstein metric over the full scale of Sobolev regularity levels $\alpha > 0$. The upper bound rate result extends that of Theorem 3.1 by Dedecker *et al.* [2] to the case where the mixing distribution is assumed to satisfy only moment conditions and is not necessarily Lebesgue absolutely continuous (a case that can be thought to correspond to $\alpha = 0$). In absence of smoothness/regularity assumptions, when $\beta \leq \frac{1}{2}$, the upper bound matches with the lower bound for every $p \geq 1$, while there is a mismatch when $\beta > \frac{1}{2}$ and $p > 1$. So, the crucial open question concerns minimax-optimal rates in the general case when the mixing distribution is regular.

References

1. Caillerie, C., Chazal, F., Dedecker, J., Michel, B.: Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.* **5**, 1394–1423 (2011)
2. Dedecker, J., Fischer, A., Michel, B.: Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electron. J. Stat.* **9**, 234–265 (2015)
3. Dedecker, J., Michel, B.: Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *J. Multivar. Anal.* **122**, 278–291 (2013)
4. Delaigle, A.: Deconvolution kernel density estimation. In: Yi, G.Y., Delaigle, A., Gustafson, P. (eds.) *Handbook of Measurement Error Models*. Chapman & Hall/CRC Press, Boca Raton (2021)
5. Rousseau, J., Scricciolo, C.: Wasserstein convergence in Bayesian deconvolution models. Technical report (2021). <https://arxiv.org/abs/2111.06846>
6. Rousseau, J., Scricciolo, C.: Wasserstein convergence in Bayesian and frequentist deconvolution models. Technical report (2023). <https://arxiv.org/abs/2309.15300>



Capturing Uncertainty in Gaussian Mixture-Based Entropy Estimation via Weighted Likelihood Bootstrap

Luca Scrucca^(✉)

Alma Mater Studiorum - Università di Bologna, Bologna, Italy
luca.scucca@unibo.it

Abstract. Entropy estimation plays a crucial role in various fields, such as information theory, statistical data science, and machine learning. However, traditional entropy estimation methods often struggle with complex data distributions. Recently, mixture-based entropy estimation has been proposed and gained attention due to its ease of use and accuracy. This work proposes a novel approach based on the weighted likelihood bootstrap to quantify the uncertainty associated with mixture-based entropy estimation. Unlike standard methods, our approach leverages the underlying mixture structure by assigning random weights to observations in a weighted likelihood bootstrap procedure, leading to more accurate uncertainty estimation. The proposed approach is illustrated by analyzing the log-returns of Brent oil prices from NYMEX for the years 2014–2023. Quantifying the inherent randomness and its uncertainty offers valuable insights into the market’s resilience, volatility patterns, and potential future trends. This study not only advances uncertainty estimation in mixture-based entropy but also provides a unique perspective on the dynamics of the oil market, with potential implications for risk management and investment strategies.

Keywords: differential entropy estimation · Gaussian mixtures · weighted likelihood bootstrap · Brent oil log-returns modeling

1 Mixture-Based Entropy Estimation

Entropy is a core concept in information theory [9] providing a measure of uncertainty or randomness in a random variable [1]. The (differential) entropy of a continuous random variable $X \in \mathbb{R}^d$ with probability density function (pdf) $f(\mathbf{x})$ is defined as

$$H(X) = -\mathbb{E}[\log f(X)] = -\int_{\mathcal{X}} f(\mathbf{x}) \log f(\mathbf{x}) \, d\mathbf{x}$$

where $\mathcal{X} = \{\mathbf{x} : f(\mathbf{x}) > 0\}$ is the support of the random variable.

Closed-form expressions for the entropy of several univariate distributions are available in the literature [3], but only few multivariate distributions admit

an analytical solution. For parametric pdfs $\{f(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, when a closed-form expression is available, using the MLE $\hat{\boldsymbol{\theta}}$ guarantees that the estimate of the entropy is both asymptotically unbiased and efficient [2, Theorem 7.3]. If we cannot assume a specific parametric distribution, an estimate of the entropy is often obtained using nonparametric estimation of $f(\mathbf{x})$ based on histogram, kernel density, or nearest neighbour estimators. This latter approach is, however, limited to low-dimensional cases, often $d \leq 2$.

Recently, an efficient mixture-based estimator has been proposed by [6] using finite mixture models (FMMs) with G components to approximate the distribution of interest, i.e.

$$f(\mathbf{x}) \approx \sum_{k=1}^G \pi_k f_k(\mathbf{x}),$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)$ are the mixing weights (s.t. $\pi_k > 0, \sum_{k=1}^G \pi_k = 1$), and $f_k(\cdot)$ is the density of the k th component of the mixture ($k = 1, \dots, G$). From a generative point of view, FMMs can be re-expressed as a hierarchical model by introducing the multinomial discrete latent variable $Z \in \{1, \dots, G\}$ and writing

$$\begin{aligned} Z &\sim \mathcal{M}(1, \boldsymbol{\pi}) \\ X \mid (Z = k) &\sim f_k(\mathbf{x}). \end{aligned}$$

Then, using the law of total expectation, the entropy can be written as

$$\begin{aligned} H(X) &= -\mathbb{E}_Z [\mathbb{E}_{X|Z} [\log f(X)]] = -\sum_{k=1}^G \pi_k \mathbb{E} [\log f(X) \mid Z = k] \\ &= -\sum_{k=1}^G \pi_k \int_{\mathcal{X}} f_k(\mathbf{x}) \log f(\mathbf{x}) \, d\mathbf{x}. \end{aligned}$$

Based on this formulation, it was shown [6] that a simple closed-formula mixture-based estimator for samples of size n can be calculated as follows:

$$\hat{H}(X) = -\frac{1}{n} \sum_{i=1}^n \log \hat{f}(\mathbf{x}_i). \quad (1)$$

Gaussian mixture models (GMMs) correspond to the case where the general FMM is written as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^G \pi_k \phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\boldsymbol{\theta} = \{\pi_1, \pi_2, \dots, \pi_{G-1}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G\}$ are the unknown parameters of the mixture, with $(\pi_1, \pi_2, \dots, \pi_G)$ the mixing weights, and $\phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ the underlying multivariate Gaussian pdf for the k th component with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The entropy estimate from (1) can then be written as

$$\hat{H}(X) = -\frac{1}{n} \sum_{i=1}^n \log \left\{ \sum_{k=1}^G \hat{\pi}_k \phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) \right\}. \quad (2)$$

2 Uncertainty Assessment via Weighted Likelihood Bootstrap

The standard nonparametric bootstrap procedure involves repeated generation of random weights from a multinomial distribution with equal probabilities, i.e. $\mathcal{M}(1; 1/n, \dots, 1/n)$. In contrast, Weighted Likelihood Bootstrap (WLB) is a resampling procedure that generate bootstrap samples by assigning random weights to observations generated from a Dirichlet over the $(n-1)$ -simplex, i.e. $(w_1, \dots, w_n) \sim \text{Dir}(\boldsymbol{\alpha})$, such that $0 \leq w_i \leq 1$ and $\sum_{i=1}^n w_i = 1$. Usually, $\boldsymbol{\alpha} = (1, 1, \dots, 1)^\top$, so values are generated uniformly over the simplex, and weights are normalized using $w_i \leftarrow w_i/\bar{w}$, where $\bar{w}$ is the average of unnormalized weights, so $\sum_{i=1}^n w_i = n$.

WLB was originally proposed by [5] as a method for approximately sampling from a posterior distribution. In its essence, WLB combines weighted likelihood estimation with bootstrap resampling using random weights as in the Bayesian bootstrap approach [7]. Recently, Weighted Bayesian Bootstrap (WBB) has been proposed by [4] for statistical and machine learning models including prior information.

WLB requires the optimization of a weighted likelihood function, which for Gaussian mixtures is given by

$$\ell^W(\boldsymbol{\theta}) = \sum_{i=1}^n w_i \log \left\{ \sum_{k=1}^G \pi_k \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

EM algorithm can be used for estimation by defining the weighted complete-data log-likelihood:

$$\ell_C^W(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^G z_{ig} w_i \log \{ \pi_k \phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \} \quad (3)$$

where $\boldsymbol{\theta} = (\pi_1, \dots, \pi_{G-1}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G)$.

The WLB procedure for a GMM($\mathcal{M}, G$), where $\mathcal{M}$ is the model determined by the selected parsimonious covariance decomposition [8, Sec. 2.2.1] and G is the number of mixture components, is the following:

1. generate $(w_1, \dots, w_n) \sim \text{Dir}(\boldsymbol{\alpha})$ and get weights as $w_i \leftarrow w_i/\bar{w}$;
2. fit a GMM($\mathcal{M}, G$) on $\{\mathbf{x}_i, w_i\}_{i=1}^n$ by maximizing (3) to get the bootstrap estimate of parameters $\hat{\boldsymbol{\theta}}_b^* = \{\hat{\pi}_k^*, \hat{\boldsymbol{\mu}}_k^*, \hat{\boldsymbol{\Sigma}}_k^*\}_{k=1}^G$;
3. compute the bootstrap estimate of entropy as

$$\hat{H}_b^*(X) = -\frac{1}{n\bar{w}} \sum_{i=1}^n w_i \log \left\{ \sum_{k=1}^G \hat{\pi}_k^* \phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k^*, \hat{\boldsymbol{\Sigma}}_k^*) \right\};$$

4. replicate steps 1–3 a large number of times, say B , to obtain the WLB distribution $(\hat{H}_1^*, \hat{H}_2^*, \dots, \hat{H}_B^*)$ as an approximate posterior sampling.

WLB credible intervals at the level $(1 - \alpha)100\%$ can be obtained by computing $[\hat{H}_{\alpha/2}^*, \hat{H}_{1-\alpha/2}^*]$, where $\hat{H}_q^*$ is the 100 q th percentile of the WLB distribution.

3 Brent Crude Oil Price Changes

Differential entropy can be applied to evaluate the randomness in time series data. For instance, in financial markets, entropy might be used for studying and understanding the movements of stock prices over time.

The price of Brent crude oil traded on the New York Mercantile Exchange (NYMEX) represents a benchmark for global oil prices. Over the past decade, it has shown strong fluctuations determined by various factors, such as geopolitical tensions, supply-demand dynamics, and macroeconomic indicators. The distribution of price changes and the degree of uncertainty over different time horizons can be used to assess the information content embedded in Brent price time series data.

Figure 1a shows the daily prices of Brent crude oil at NYMEX. After peaking in mid-2014, prices declined until early 2016, then slowly increased and stabilized. At the beginning of 2020, corresponding to the spread of the COVID-19 pandemic, the price of Brent reached its lowest value, falling below 25 USD per barrel. Thereafter, a rapid increase was observed in the following years, with a more stable trend observed in 2023. These behaviors are reflected in the plot of the log-returns shown in Fig. 1b, which exhibits strong fluctuations in the years 2015–16 and 2020, while relatively more stable patterns are shown in the years 2014, 2017–2019 and 2023.

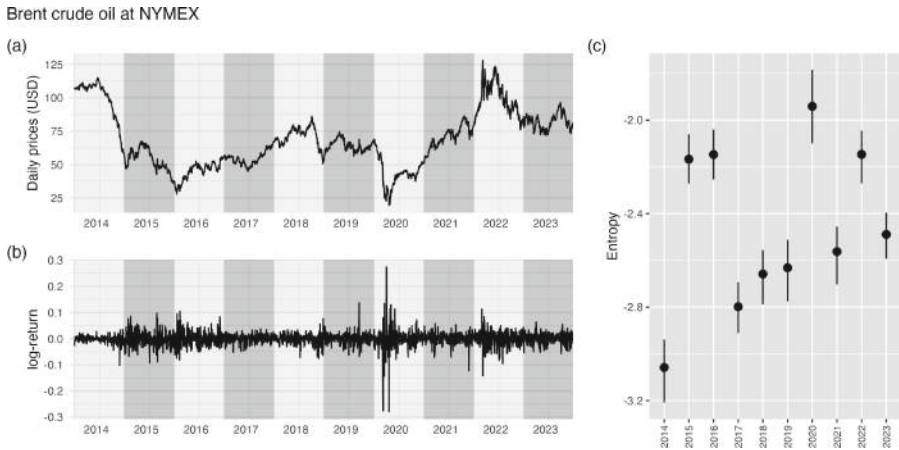


Fig. 1. Brent crude oil prices (USD) at NYMEX from 2014 to 2023 (a) and corresponding log-return values (b). Yearly entropy estimates and 95% credible intervals obtained using the weighted likelihood bootstrap approach.

Using the distribution of log-returns for each year from 2014 to 2023, we fitted GMMs with unconstrained variances across components and number of mixture components selected by BIC. Entropy was then calculated using (2) and the corresponding uncertainty assessed following the approach discussed in Sect. 2.

The yearly entropy estimates with 95% credible intervals obtained via WLB are shown in Fig. 1c. We observe that the lowest entropy occurs in 2014, while the highest value is seen in 2020. The years 2015–2016 and 2022 also exhibited high volatility in Brent oil prices, whereas the remaining years appear more stable.

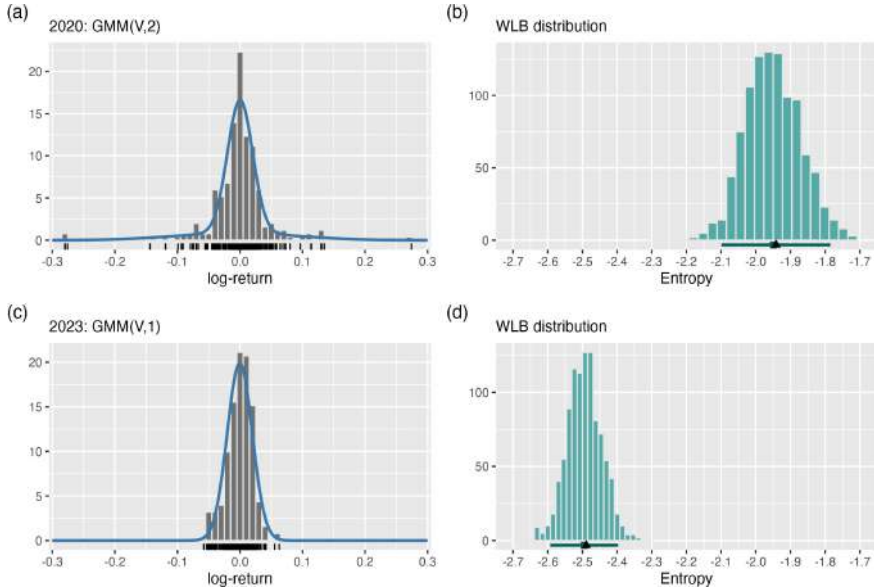


Fig. 2. Distribution of log-returns for Brent crude oil at NYMEX for the years 2020 (a) and 2023 (c), with the corresponding bootstrap distribution for the entropy obtained via WLB (b and d). The latter graphs also show the entropy estimate (black triangle) and the 95% credible intervals (green bottom line). (Color figure online)

Finally, the left panels of Fig. 2 show the distribution of log-returns for the years 2020 (a) and 2023 (b) with the respective GMM density estimates. The right panels of Fig. 2 report the corresponding bootstrap distributions obtained via WLB, showing the entropy estimates and the 95% credible intervals.

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. John Wiley & Sons, Hoboken (2006)
2. Kay, S.M.: Fundamentals of Statistical Signal Processing. Prentice Hall, Upper Saddle River (1993)
3. Michalowicz, J.V., Nichols, J.M., Bucholtz, F.: Handbook of Differential Entropy. Chapman & Hall/CRC, Boca Raton (2014)
4. Newton, M.A., Polson, N.G., Xu, J.: Weighted bayesian bootstrap for scalable posterior distributions. *Can. J. Stat.* **49**(2), 421–437 (2021). <https://doi.org/10.1002/cjs.11570>

5. Newton, M.A., Raftery, A.E.: Approximate bayesian inference with the weighted likelihood bootstrap (with discussion). *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)* **56**, 3–48 (1994)
6. Robin, S., Scrucca, L.: Mixture-based estimation of entropy. *Comput. Stat. Data Anal.* **177**, 107582 (2023). <https://doi.org/10.1016/j.csda.2022.107582>
7. Rubin, D.B.: The bayesian bootstrap. *Ann. Stat.* **9**(1), 130–134 (1981)
8. Scrucca, L., Fraley, C., Murphy, T.B., Raftery, A.E.: Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman & Hall/CRC, Boca Raton (2023). <https://doi.org/10.1201/9781003277965>
9. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)



A New Approach for Text Bi-Clustering

Ilaria Bombelli¹, Domenica Fioredistella Iezzi², and Emiliano Seri²(✉)

¹ Istat, Rome, Italy

`ilaria.bombelli@istat.it`

² Department of Enterprise Engineering, University of Rome Tor Vergata,
Rome, Italy

`{stella.iezzi,emiliano.seri}@uniroma2.it`

Abstract. In this study, we introduce the Double Spherical K-Means (DSKM) clustering method for text data. A novel approach for simultaneous clustering of terms and documents. Leveraging the strengths of k-means, double k-means, and spherical k-means, DSKM addresses the challenges of high dimensionality, noise, and sparsity inherent in text analysis. We apply DSKM to the corpus of US presidential inaugural addresses, spanning from George Washington in 1789 to Joe Biden in 2021. Our analysis reveals distinct clusters of words and documents that correspond to significant historical themes and periods, showcasing the method's ability to facilitate a deeper understanding of the data. Our findings demonstrate DSKM's efficacy in uncovering underlying patterns in textual data.

Keywords: Text data · Co-clustering · Double spherical k-means

1 Introduction

Text data has become increasingly important in recent years due to the massive amount of data being generated in various forms, including social networks, patient records, healthcare insurance data, and news outlets. Text clustering is a set of techniques and models that utilizes algorithms to group similar documents from a corpus without any predetermined categories or labels [Iezzi, 2012]. The goal is to reveal undiscovered patterns, topics, or clusters in a collection of text documents, which in turn facilitates the exploration and organization of large text datasets based on their inherent similarities. Among the most used clustering algorithms, k-means and its variants quickly classify large and unstructured data. These algorithms groups data points into clusters based on their similarity and offers various benefits such as: ease of interpretation, scalability, and guaranteed convergence. Despite its advantages, k-means clustering has several limits. These include the need to pre-determine the number of clusters, sensitivity to outliers, and the possibility of getting stuck in local minima. A commonly used method to calculate text similarity is based on the distance between text vectors in vector space [Salton and Buckley 1988], but the Euclidean distance

in fact is not suitable for text data, since long documents characterized by high term weights are over-represented [Hornik et al. 2012]. One of the most useful distances in text clustering is the cosine distance. Cosine distance is a measure of the dissimilarity between two vectors based on the cosine angle between them [Maher and Joshi 2016]. In this paper, we present an original method for simultaneous partitioning of terms and documents called Double Spherical K-Means (DSKM) clustering. Co-clustering, or clustering of both rows and columns of a Term-Document-Matrix [Celardo et al. 2016], helps in classifying terms into clusters based on the complete set of documents and vice versa. The assignments of documents to clusters are based on the entire set of terms. We apply the method to the US presidential inaugural addresses data, available in the R package *quanteda* [Benoit et al. 2018].

2 Methodology

Co-clustering methodologies proves particularly advantageous in scenarios marked by sparsity, noise, and high dimensionality. By exploiting the duality between row and column clusters, co-clustering methods foster robustness against the challenges posed by large, sparse, and non-negative matrices, frequently encountered in textual data analysis. Throughout the co-clustering process, the relationships between row and column clusters are intricately intertwined, facilitating a more comprehensive understanding of the underlying data structure. The proposed DSKM methodology integrates established techniques such as k-means clustering, double k-means clustering (DKM) [Vichi2001], and spherical k-means (SKM) [Hornik et al. 2012, [Dhillon and Modha, 2001], tailored specifically for textual data analysis. By unifying these methodologies, DSKM offers a versatile approach with broad applicability in various objectives, including but not limited to topic modeling, document clustering, sentiment analysis, information retrieval, and document summarization.

Consider a word by-document matrix $\mathbf{X} = \{x_{ij} : i = 1, \dots, N, j = 1, \dots, J\}$ where i indicates words and j indicates documents. The optimization problem of DSKM is stated below:

$$\text{minimize } J_{DSKM} = \|\mathbf{X} - \mathbf{UCV}'\|^2 \quad (1)$$

s.t.

$$\begin{aligned} \sum_{k=1}^K u_{ik} &= 1 & \forall i = 1, \dots, N \\ \sum_{q=1}^Q v_{jq} &= 1 & \forall j = 1, \dots, J \\ \|\mathbf{c}_k\| &= 1 & \forall k = 1, \dots, K, \end{aligned} \quad (2)$$

where: $\mathbf{X}$ is the $(N \times J)$ data matrix; $\mathbf{U}$ is the $(N \times K)$ membership matrix, $u_{ik} = 1$ if the unit i belongs to cluster k , 0 otherwise and satisfying $\sum_{k=1}^K u_{ik} =$

$1 \forall i = 1, \dots, N$; $\mathbf{V}$ is the $(J \times Q)$ membership matrix, $v_{jq} = 1$ if the variable j belongs to cluster q , 0 otherwise and satisfying $\sum_{q=1}^Q v_{jq} = 1 \forall j = 1, \dots, J$; $\mathbf{C}$ is the $(K \times Q)$ centroids matrix, with the property that centroids related to cluster of units are unit-vectors.

In order to optimize the constrained problem, the following algorithm has been used:

0. Initialization: Randomly generate matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ from a uniform distribution and make them row-stochastic.
1. Normalize data matrix: Normalize the data matrix by row, i.e. $\mathbf{Z} = \{z_{ij}, i = 1, \dots, N, j = 1, \dots, J\}$ s.t. $\|\mathbf{z}_i\| = 1 \forall i = 1, \dots, N$
2. Compute centroids: Given $\hat{\mathbf{U}}, \hat{\mathbf{V}}$, the updating formula for the kq -th centroid is the following:

$$m_{kq} = \frac{\sum_{i=1}^N \sum_{j=1}^J u_{ik} v_{jq} x_{ij}}{\sum_{i=1}^N \sum_{j=1}^J u_{ik} v_{jq}} \quad (3)$$

3. Normalize centroids: Given $\hat{\mathbf{M}}$ the updating formula for the k -th normalized centroid for the q -th cluster is the following:

$$\mathbf{c}_k = \frac{\mathbf{m}_k}{\|\mathbf{m}_k\|} \quad (4)$$

4. Update $\mathbf{U}$: Given $\hat{\mathbf{V}}, \hat{\mathbf{C}}$, compute all the cosine similarities between the observation $\mathbf{z}_i$ of matrix $\mathbf{Z}$ and the centroid $\mathbf{c}_k \forall k = 1, \dots, K$ of matrix $\mathbf{C}$ (properly transformed into a $K \times J$ matrix). Then assign observation i to the cluster corresponding to the closest centroid.
5. Compute and normalize centroids: as in step 2 and 3, respectively.
6. Update $\mathbf{V}$: Given $\hat{\mathbf{U}}, \hat{\mathbf{C}}$, compute all the cosine similarities between the observation $\mathbf{z}_j$ of matrix $\mathbf{Z}$ (properly normalized by column) and the centroid $\mathbf{c}_q \forall q = 1, \dots, Q$ of matrix $\mathbf{C}$ (properly transformed into a $N \times Q$ matrix). Then, assign observation j to the cluster corresponding to the closest centroid.
7. Compute the objective function: Compute the objective function.
8. Check convergence: Repeat steps 2–7 until convergence is reached. In other words, stop when $|J_{DSKM}^{(t)} - J_{DSKM}^{(t-1)}| < \epsilon$.

In order to avoid local minima, the algorithm is suggested to run from multiple random starts (initial randomly generated membership matrices).

3 Application

We applied the DSKM method on a dataset containing the US presidential inaugural addresses. These speeches were delivered by all US presidents from George Washington in 1789 to Joe Biden in 2021. We aim to investigate if any shared topics can be identified in clusters of speeches delivered by different presidents. The initial corpus comprises 59 documents, with a total of 151,536 tokens, 9,442 types, and 4,218 hapaxes. On average, each document contains 770.4 types and

2,568 tokens. We calculated the type/token ratio, which is 0.0623, and the hapax percentage, which is 0.4467. These values suggest a high lexical variety, justifying the computational analysis of the corpus. Specifically, we used the DSKM to cluster both the documents and the words simultaneously. During the first stage of our analysis, we focused on cleaning and pre-processing the data. This step is crucial for text analysis. To achieve this, we used tokenization to break the text into smaller units and removed numbers, punctuation, symbols, and separators. We also applied lemmatization to bring each token to its basic form. To create a document-term matrix, we removed stop-words and trimmed the data by removing all words with a frequency less than or equal to 10. We utilized the TF-IDF weighting scheme, which evaluates the significance of a word for a text or class of texts [Liang and Niu 2022]. To avoid any bias towards longer documents, which may have a higher term count even if the term is not particularly important, we used the normalized term frequency. After pre-processing, we ended up with a final data matrix consisting of 59 documents and 1,206 types, this matrix results to be sparse at 67.73%.

3.1 Double Spherical k-means

To determine the best number of clusters for both terms and documents, we used three commonly used methods - elbow, silhouette, and gap statistic. According to all three methods, the optimal number of clusters for words is 3. However, they provided different results for the documents, with silhouette indicating 2 clusters, elbow indicating 3 clusters, and gap statistic indicating 1 cluster. For better interpretability, we decided to use 2 clusters. [Fraley and Raftery 1998].

Figure 1 shows the clusters of documents identified by the DSKM method. It can be clearly seen how the two groups have a chronological distinction between them. Upon analyzing the most frequently used words within each cluster, we were able to identify distinct themes for each of the three clusters. The first cluster groups together words related to patriotism and change, which we have named the “American dream” cluster. The second cluster consists of words related to law and war, which we have grouped as the “law and wartime” cluster. Finally, the third cluster is made up of words associated with politics and economics, which we have named the “politics and economics” cluster.

Figure 2 displays the 30 most frequently occurring words within two clusters of documents, categorized by their respective word clusters. The first. and second clusters of words are distinctly distributed among the first and second clusters of documents, respectively. The third cluster of words is less frequent and only appears in the top 30 words of the documents.

There are distinct differences between the language and topics covered by the first and second groups of presidents. The first group uses words related to the Revolutionary War, the Texas Revolution, and the abolition of slavery. On the other hand, the second group uses words that refer to the Second World War, the Cold War, the transition to the 21st century, and patriotism.

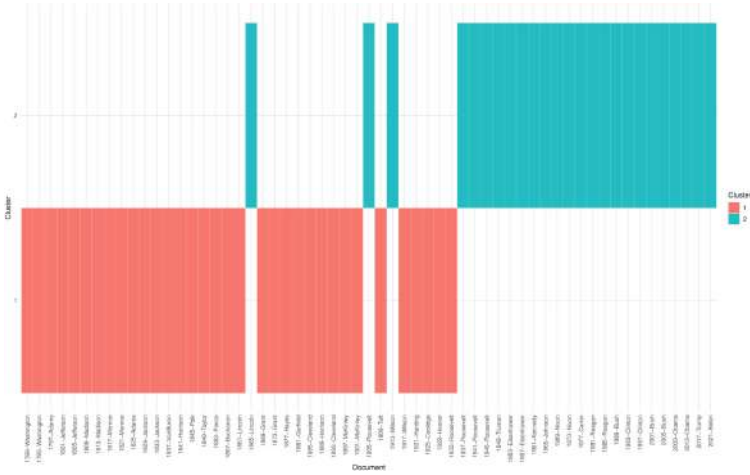


Fig. 1. Chronologically ordered US presidential inaugural addresses, colored by cluster of documents

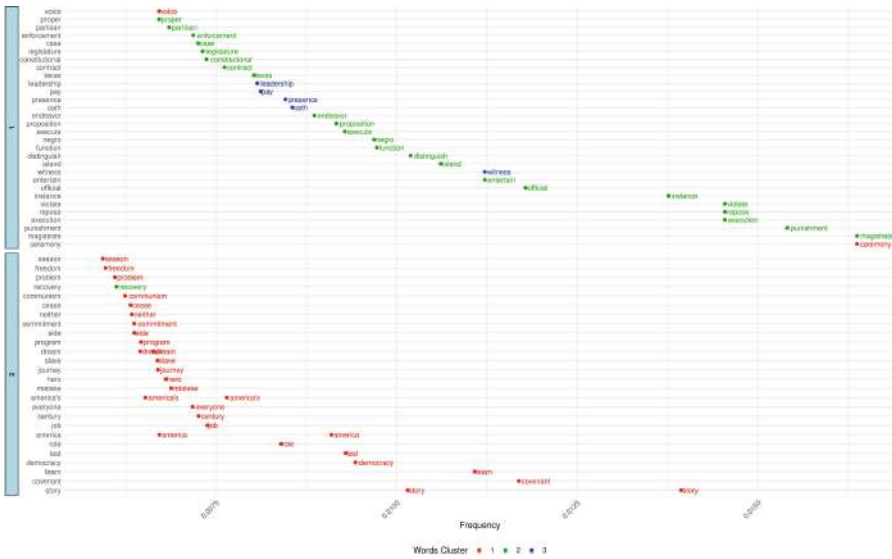


Fig. 2. Top 30 words in each cluster of documents, colored by cluster of words

4 Final Consideration

We developed a method called DSKM which can simultaneously partition terms and documents in text data. DSKM integrates the principles of k-means, double k-means, and spherical k-means clustering, creating a robust framework that adeptly addresses the challenges of high dimensionality, sparsity, and inherent

noise in text data. When we applied this method on the US presidential inaugural addresses data, it not only identified distinct clusters of topics that have evolved over time but also offered new insights into the changing focuses of presidential discourse. This research highlights the utility of DSKM in extracting meaningful patterns from complex text corpora, making it a valuable tool across various domains.

References

- Iezzi, D.F.: Centrality measures for text clustering. *Commun. Stat. Theory Methods* **41**(16–17), 3179–3197 (2012)
- Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**(5), 513–523 (1988). ISSN 0306-4573. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0). <https://www.sciencedirect.com/science/article/pii/S0306457388900210>
- Hornik, K., Feinerer, I., Kober, M., Buchta, C.: Spherical k-means clustering. *J. Stat. Softw.* **50**, 1–22 (2012). <https://doi.org/10.18637/jss.v050.i10>
- Maher, K., Joshi, M.S.: Effectiveness of different similarity measures for text classification and clustering. *Int. J. Comput. Sci. Inf. Technol.* **7**(4), 1715–1720 (2016)
- Celardo, L., Iezzi, D.F., Vichi, M.: Multi-mode partitioning for text clustering to reduce dimensionality and noises. In: *JADT 2016-Statistical Analysis of Textual Data*, vol. 1, pp. 181–192. Presses de FacImprimeur (2016)
- Benoit, K., et al.: quanteda: an r package for the quantitative analysis of textual data. *J. Open Source Softw.* **3**(30), 774 (2018). <https://doi.org/10.21105/joss.00774>. <https://quanteda.io>
- Vichi, M.: Double k-means clustering for simultaneous classification of objects and variables. In: Borra, S., Rocci, R., Vichi, M., Schader, M. (eds.) *Advances in Classification and Data Analysis*, pp. 43–52. Springer, Heidelberg (2001)
- Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **42**, 143–175 (2001). <https://doi.org/10.1023/A:1007612920971>
- Liang, M., Niu, T.: Research on text classification techniques based on improved tf-idf algorithm and lstm inputs. *Procedia Comput. Sci.* **208**, 460–470 (2022). ISSN 1877-0509. <https://doi.org/10.1016/j.procs.2022.10.064>. <https://www.sciencedirect.com/science/article/pii/S1877050922015058>
- Fraley, C., Raftery, A.E.: How many clusters? which clustering method? answers via model-based cluster analysis. *Comput. J.* **41**(8), 578–588 (1998). ISSN 0010-4620. <https://doi.org/10.1093/comjnl/41.8.578>



Language of Bankruptcy: Analysis of Word Sequences and Word Contexts

Andrea Simonetti^(✉) and Rodolfo Damiano

University of Palermo, Palermo, Italy
andrea.simonetti01@unipa.it

Abstract. Academics and practitioners searched for reliable indicators of companies' failure focusing only on quantitative data such as financial ratios and market variables. However, recent literature aims to quantify textual information of financial reports studying features such as topics and words' co-occurrences, confirming their usefulness in predicting company bankruptcy. In this work, we propose a new approach to analysing texts that focuses on sentences interpreted as ordered sequences of words. We propose a new approach, based on Language Model, to predict the company's bankruptcy that was released in the next year. Given the high predictive power of the model, we investigate the sentences of texts to gain insights into how failing companies' language differs from the non-failing one. Our approach allows us to move away from fixed word-lists, exploring linguistic features to understand how a word is used in different contexts. The results of our analysis lead us to observe that the concept of bankruptcy can take on different meanings arising from the different legitimisation strategies that companies facing bankruptcy may use.

Keywords: Language Models · Bankruptcy prediction · word co-occurrences

1 Introduction

Corporate bankruptcy is one of the most important credit risk factors and attracts the attention of creditors and investors. Therefore, an accurate bankruptcy forecasting model is essential for practitioners, regulators, and academic researchers who can use it to supervise the financial health of individual institutions, contain systemic risks and predict default probability to price corporate debt [11]. First, accounting and finance studies monitored credit risk and bankruptcy prediction focusing only on quantitative data such as financial ratios and market variables [2]. Recent Business Failure Prediction (BFP) literature focused on incorporating textual content extracted from annual disclosures to the respective regulatory authorities [3]. Recent works on text analysis of financial reports aim to quantify textual information as significant variables to predict future company financial performance [4, 8]. Some of these studies confirmed text data's explanatory power and usefulness in predictive tasks, such as discriminating between bankruptcy and non-bankruptcy companies. In other words, these

studies seem to prove that a “bankrupt language” exists. However, most of these works on textual analysis rely on descriptive text statistics techniques such as sentiment word count, length, spelling errors, tones and readability [6, 7]. These approaches present some limitations related to the pitfalls of investigating the complexity and fuzziness of natural language.

2 Data Collection

Our analysis investigates USA companies’ 10-K filings, specifically on MD&A section. These reports are required by the Securities and Exchange Commission (SEC), and are publicly available. To create our 10-Ks’ sample, we relied on two different sources of data: the UCLA-LoPucki Bankruptcy Research Database (UBRD)¹ and the Loughran and McDonald texts repository². We matched companies by SIC codes (Industrial Sector), year of reports and Total Assets (control company within 10% of total assets of the match) as in previous works [4]. Then, we obtained a sample of 82 bankruptcies and 82 non-bankruptcies. For each of these bankrupt companies, we collected the 10-Ks related to the year before failure, and 10-Ks were collected for the matched healthy companies during the same period.

3 Methods

3.1 Language Model

Statistical Language Model is an important method widely used in many applications in computational linguistics to face tasks such as speech recognition, machine translation and natural language generation. It uses probabilities to estimate how likely any given sequence of words belongs to a language. It attempts to reflect the frequency with which each sequence of length t occurs as a sentence in natural text. Specifically, N -gram models decompose the probability of a sequence of words into conditional probabilities of each word given the previous context by calculating n -gram probabilities, where an n -gram is defined as a sequence of n words. In these models, the Markov assumption has a central role in defining the time dependencies of the sequence of words. It means that the probability of observing a word only depends on a fixed number of previous words. So, the probability of a word only depends on a fixed number n of previous words.

$$P(w_t|w_1, \dots, w_{t-1}) = P(w_t|w_{t-n+1}, \dots, w_{t-1}) \quad (1)$$

¹ LoPucki, L. M. (2015). UCLA-LoPucki bankruptcy research database user’s manual. Unpublished manuscript. <http://lopucki.law.ucla.edu/index.htm>, retrieved October 01, 2020.

² <https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/>.

In particular, our focus is on assigning probabilities to sentences. Considering the joint probability of a sentence of t words as follows

$$P(w_1, \dots, w_t) = P(w_1)P(w_2|w_1) \dots P(w_t|w_1, \dots, w_{t-1}) = \prod_i^t P(w_i|w_{j < i}) \quad (2)$$

These models need techniques to face the problem of assigning non-zero probabilities to sequences that may never be seen in the training corpus. The solution is to “smooth” the language model to move some probability towards unknown n -grams. One of the most effective smoothing methods is the Kneser-Ney Smoothing [5] because of its use of absolute-discounting interpolation, which consists in subtracting a fixed value from the probability’s lower order terms to leave out n -grams with lower frequencies.

3.2 Statistical Test on Word Context

We also investigate language patterns in the narration of bankruptcy and healthy companies and gain insights into how failing companies’ language differs from non-failing ones. We assume that bankruptcy companies should use some words in a significantly different way than healthy ones. In particular, we focus on understanding how words are used in the “fail” or “health” context. In this respect, we can assign a specific word to the bankruptcy or healthy category by performing a statistical test. So, for each company, we test if its words have context more similar to the bankruptcy or healthy cases. We represent words in a vector space where each dimension corresponds to a word in the Vocabulary and we counts its co-occurrences with other words within a window of length 2, as the order of the Markov model described before. So, for each word we have a vector representation that describes its word-context. Once we obtain the word vector representation for every word for a specific company, we compute the profile of the same word in the population of bankrupt documents and healthy documents. Then, for each word, we collect two lists of cosine similarities from the pairwise comparison between the document and those related to the two corpora. So, we have two lists of cosine similarity scores as follow: $S_{k,b} = \{x_1, \dots, x_i, \dots, x_n\}$ that represents the similarity scores for the k -th word between the document to test and the n documents in the bankruptcy corpora, and $S_{k,h} = \{y_1, \dots, y_j, \dots, y_n\}$ that represents the similarity scores for the k -th word between the document to test and the n documents in the healthy corpora. Then, we perform the Bootstrap t-test [12] to compare the means of these two independent samples of word-context similarities, setting the significant level α equal to 0.01. Then, if the difference is statistically significant we assign a label “negative” or “positive” if the mean of the similarities scores is higher in the bankruptcy or healthy corpora, respectively.

4 Results

4.1 Classification Performance

We face the task of prediction one year ahead the possible failure of a company analysing the language’s style of the text. To this end, we create two corpora: one related to companies that went bankrupt the year following one of the analyzed MD&A and the other to companies that remained solvent the next year. Then, we train two Language Models, one for each corpus. The trained models estimate two probabilities, respectively, for all sentences of an unseen document, assigning the probability of how likely a sentence comes from the corpus of bankruptcy and the probability that it comes from the healthy corpus. We apply a leave-one-out cross-validation to the 82 pairs of matched companies’ documents to evaluate the model’s performance. So, we have two documents in the test set. For each document on the test set and all sentences in the document, we collect the logarithm of the probabilities predicted by the two models (one from the corpus of “will-fail” companies and one from the corpus of “healthy” companies) separately. Then, we compute a score for a document d_j according to the following formula:

$$score_M(d_j) = exp\{\frac{1}{N} \sum_{i=1}^N p_M(s_i)\} \quad (3)$$

where N represents the number of sentences in the document d_j , $M \in \{F, H\}$ identifies from which model the log probabilities $p_M(\cdot)$ come from (“will-fail” or “healthy”). So, we collect 82 pairs of scores and we assign a prediction label, H or F , according to the highest score’s model. In Table 1, we compare our model with the others proposed in the literature with the goal of bankruptcy prediction of public companies in the U.S., including text variables retrieved from the MD&A section of 10-K reports. As shown in the Table 1, our model’s accuracy is the highest in the literature concerning bankruptcy prediction through the analysis of MD&A texts.

Table 1. Prediction of bankruptcy: accuracy performance comparison

Method	Accuracy	Variables
Language Model	90%	Textual
Support Vector Machine [4]	83%	Textual & Altman [1]
Deep Learning Model [10]	71%	Textual, Accounting & Market

4.2 Language of Bankruptcy

Then, we focus on linguistic features. In particular, we want to understand how a word’s context influences its association with failure or non-failure. For each

company, we can divide all sentences into two categories, which we call “negative” and “positive”, according to the highest log-probability between the two different trained models. Then, we select the corpora made by negative sentences of bankrupt companies and the other one by the positive sentences of healthy companies. Then, we use the sentiment word list of Loughran and McDonald [9]. In particular, we find that some words tagged as “positive” could have a different meaning in different contexts. Table 2 below lists some representative sentences for each corpora. These results show the efficacy of the Language Model in managing the local context of words and understanding where the sentences are more likely to come from.

Table 2. Different use of positive words in sentences of the two corpora

Stemmed words	Negative sentences & bankrupt companies	Positive sentences & healthy companies
advantag	We may also be prevented from taking advantage of business opportunities that arise because of the limitations imposed on us by such restrictive covenants	Subject to financing alternatives, we may also increase our capital expenditures significantly to take advantage of opportunities we consider to be attractive
satisfi	Our inability to generate sufficient cash flow to satisfy our debt obligations, including obligations under the notes, or to obtain alternative financing, could materially and adversely affect our business, financial condition, results of operations and prospects	Allocate the transaction price to the performance obligation in the contract, and recognize revenue as the entity satisfies performance obligations
success	We provide no assurances that we will be able to successfully consummate the Restructuring or other alternatives to restructure our existing indebtedness, in which case we may need to restructure under the Bankruptcy Code	Our production continues to grow through drilling success as we place new wells on production and through additions from acquisitions partially offset by the natural decline of our natural gas and oil reserves through production and asset sales
suffici	Additional sources of liquidity in the future as a result of our inability to generate sufficient cash flow from operations to service our long-term capital needs	Cash flow from operations and available borrowings under our revolving credit facility will be sufficient to meet our liquidity needs in the coming twelve months

5 Conclusions

Our work contributes to previous literature investigating bankruptcy using textual variables by confirming past findings on the incremental information of textual data and demonstrate that a bankruptcy language exists. The results show

that there is neither a single dictionary indicative of bankruptcy nor a unique polarization for a specific word. So, our proposal overcomes the issue of word sense disambiguation inherent in the classic dictionary approach. Moreover, we demonstrated that a bankrupt language category exists, and it is characterized by the contexts in which words are used. The proposed approach allows us to understand the linguistic style of companies' reports and to select specific "red flag patterns" (one for each report) regarding negative sentences or words with a bankruptcy profile. Finally, further developments could concern implementing this approach to investigate other parts of the report (like the accounting policy section) or different kinds of reports (such as sustainability and integrated reports) and phenomena (such as fraud and integrated thinking), including the accounting and financial information. Despite the encouraging results, this work is not without limitations. In particular, one of the main limitations is the sample size. The improvements could be expanding our sample and testing our approach using an unbalanced dataset.

References

1. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finan.* **23**(4), 589–609 (1968)
2. Altman, E.I., Iwanicz-Drozdzowska, M., Laitinen, E.K., Suvas, A.: Financial distress prediction in an international context: a review and empirical analysis of altman's z-score model. *J. Int. Finan. Manag. Account.* **28**(2), 131–171 (2017)
3. Borchert, P., Coussement, K., De Caigny, A., De Weerd, J.: Extending business failure prediction models with textual website content using deep learning. *Eur. J. Oper. Res.* (2022)
4. Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P.: Making words work: using financial text as a predictor of financial events. *Decis. Supp. Syst.* **50**(1), 164–175 (2010)
5. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: 1995 International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 181–184. IEEE (1995)
6. Li, F.: Annual report readability, current earnings, and earnings persistence. *J. Account. Econ.* **45**(2–3), 221–247 (2008)
7. Li, F., et al.: Textual analysis of corporate disclosures: a survey of the literature. *J. Account. Lit.* **29**(1), 143–165 (2010)
8. Lopatta, K., Gloger, M.A., Jaeschke, R.: Can language predict bankruptcy? the explanatory power of tone in 10-k filings. *Account. Perspect.* **16**(4), 315–343 (2017)
9. Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *J. Finan.* **66**(1), 35–65 (2011)
10. Mai, F., Tian, S., Lee, C., Ma, L.: Deep learning models for bankruptcy prediction using textual disclosures. *Eur. J. Oper. Res.* **274**(2), 743–758 (2019)
11. Schönbucher, P.J.: Credit Derivatives Pricing Models: Models, Pricing and Implementation. John Wiley & Sons, Hoboken (2003)
12. Tibshirani, R.J., Efron, B.: An introduction to the bootstrap. *Monogr. Stat. Appl. Probabil.* **57**, 1–436 (1993)



Modelling Framework to Evaluate the Phenological Development of Vine and Potential Impact of Frost: The Case of Prosecco DOC

Anna Simonetto¹(✉), Eugenio Brentari², and Alfonso Piscitelli³

¹ Department of Civil, Environmental, Architectural Engineering and Mathematics (DICATAM), University of Brescia, Brescia, Italy

anna.simonetto@unibs.it

² Department of Economics and Management (DEM), University of Brescia, Brescia, Italy

³ Department of Agriculture, University of Naples Federico II, Naples, Italy

Abstract. The goal of this study is to develop a statistical modelling framework to investigate the potential impact of climate change on grapevine phenology curves and evaluate the impact of potential frosts on grape production. It is expected to develop a regression model to predict the phenological stage of the vine based on meteorological conditions and vineyard characteristics (e.g., orography and exposure). The second component of the framework is an impact model that investigates the potential impacts of frost based on the intensity of the frost and the phenological stage of the plant. The model will be calibrated using historical data on budburst and production from three production entities of the ‘Consorzio Prosecco DOC’ in Italy. The modelling framework addresses the need to support wine operators in dealing with the effects of climate change on production. Since frost climatic conditions and vine phenology are linked to seasonal temperature development, they are subject to long-term climate changes. Therefore, in the future, the risk of frost damage will depend not only on changes in the frequency and occurrence of frost days but also on modifications to vine phenology.

Keywords: Phenological curves · climate change · impact model · budding

1 Introduction

The wine sector in Italy represents recognized excellence worldwide and is one of the agricultural production sectors with the greatest marginality [1]. One of the threats that the sector faces is climate change, which leads to changes in production cycles and affects the quality and quantity of production.

In the last decade, mild winters have been observed, leading to an early onset of plant budding. If vine buds start to sprout prematurely, the plant may become more vulnerable to spring frost [2–4]. Cold temperatures can heavily impact plant development and the quality and quantity of grapes produced, significantly compromising production for that

agricultural season [5–7]. The temperature conditions necessary for freezing damage to plant tissues and the intensity of the effects depend on the phenological growth stage reached. The effect of frosts is generally more intense as the plant's phenological stage grows. This is because more mature buds contain a greater amount of water, so even with less severe temperatures, the formation of ice crystals that damage tissues is possible. The young tissues of the grape with high water content are severely damaged as soon as they freeze. Damaged tissues quickly lose their turgidity, become completely dark, and soak with water. Although frost damage can occur on closed buds, the budding period is the most susceptible time [8–11]. Evidence suggests that various varieties differ in critical temperatures at the same phenological stage and therefore in their susceptibility to frost damage [12]. For an estimate of the risk of late frosts in future climate scenarios, two pieces of information are needed: the budding period and the air temperature conditions during and after budding.

The objective of this study is to develop a statistical modelling framework to evaluate the potential impact of climate change on phenology curves and consequently assess the impact of potential frosts on grape production in the Prosecco DOC area.

2 Materials and Methods

The first module of the methodological framework consists of the predictive model of the phenological curve of the early stages of vine development (up to leaf development) based on meteorological conditions and environmental characteristics (orography, slope, and exposure).

For the development of this model, historical data on budding curves from 2010 to 2023 were collected in three production areas of the Prosecco DOC Consortium, along with the respective time series of meteorological conditions. The analysis focuses on the Glera cultivar, which represents the basic cultivar in the Prosecco DOC regulations. Phenological observations were carried out from the moment of vegetative awakening until the beginning of summer. The BBCH scale was used for the surveys [13].

Meteorological data related to air temperature, precipitation, wind, and air humidity were collected through the reference stations Centro Meteo ARPAV and integrated, where necessary, with two other data sources (ilmeteo.it and Visualcorssing). Different indicators were considered from the raw data, including minimum temperature, dew point, temperature difference between maximum and minimum, daily average temperature, degree days with different development thresholds, hours of cold and heat stress. External factors, such as the characteristics of each analyzed vineyard regarding orography, vineyard exposure, and slope, were also included in the model. These data will be analyzed through a linear regression model in order to predict the phenological stage of the vine (from budding to leaf development).

The second module of the modeling framework involves the evaluation, based on literature results [5] and collected historical experimental data, of the effect of frosts at different moments of the vine's phenological curve. As there is no shared definition of frost, three types of frost are considered: severe frost (rigid temperature for a long time period), moderate frost (rigid temperature but for a short period or less rigid temperature but for a long period), and mild frost (not very rigid temperature for a short period). At

the end of this phase, an impact model will be defined to link the phenological stage (BBCH degrees) and frost severity with the expected impacts on the plant and grape production.

The last module of the framework involves scenario analysis. Nine evaluation scenarios will be defined by combining three elements: i) phenological curve (based on the climatic scenario); ii) severity of frost (severe, moderate, mild); iii) timing of frost occurrence (between late winter and early spring). For each of the 9 scenarios, the stage of the vine's phenological curve at which the frost occurs and the consequent impact on the plant (in terms of quality and/or quantity of production) will be estimated.

3 Conclusion

The impact of climate change on agricultural production is a serious threat to the economic sustainability of the sector. It is therefore necessary to provide industry operators with quantitative tools capable of supporting decisions to mitigate the effects of these changes on their production processes. In viticulture, one of the most significant risks is the loss of production due to frosts. For frost damage to occur, the synchrony of early phenological development of plants with the occurrence of low-temperature events is crucial. Therefore, in the future, the risk of frost damage will depend not only on changes in the frequency and occurrence of frost days but also on the modification of vine phenology. This study will assess the potential impacts of frosts based on climate change scenarios in a productive area of particular national relevance such as the Prosecco DOC Consortium. This knowledge will support producers in deciding on strategies to protect themselves from the risk of such impacts.

References

1. ISTAT: Andamento dell'economia agricola - anno 2022 (2023)
2. Dinu, D.G., et al.: Climate change impacts on plant phenology: grapevine (*Vitis vinifera*) bud break in wintertime in Southern Italy. *Foods* **10**(11), 2769 (2021)
3. Biasi, R., Brunori, E., Ferrara, C., Salvati, L.: Assessing impacts of climate change on phenology and quality traits of *Vitis vinifera* L.: the contribution of local knowledge. *Plants* **8**(5) (2019)
4. Cogato, A., Meggio, F., Pirotti, F., Cristante, A., Marinello, F.: Analysis and impact of recent climate trends on grape composition in north-east Italy. *BIO Web Conf.* **13**(04014) (2019)
5. Meier, M., Fuhrer, J., Holzkämper, A.: Changing risk of spring frost damage in grapevines due to climate change? A case study in the Swiss Rhone Valley. *Int. J. Biometeorol.* **62**(6), 991–1002 (2018)
6. Frioni, T., et al.: Impact of spring freeze on yield, vine performance and fruit quality of *Vitis* interspecific hybrid Marquette. *Sci. Hortic.* **19**, 302–309 (2017)
7. Meggio, F.: The interplay between grape ripening and weather anomalies in Northern Italy – a modelling exercise. *OENO One* **56**(2) (2022)
8. Persico, M.J., Smith, D.E., Centinari, M.: Delaying budbreak to reduce freeze damage: seasonal vine performance and wine composition in two *Vitis vinifera* cultivars. *Am. J. Enol. Vitic.* **72**(4), 346–357 (2021)

9. Tomasi, D., Lovat, L.: La risposta fenologica della vite ai cambiamenti climatici. Il “proxy” di Conegliano, Consiglio per la Ricerca e la sperimentazione in Agricoltura Centro di ricerca per la viticoltura (2010)
10. Trought, M.C.T., Howell, G.S., Cherry, N.: Practical considerations for reducing frost damage in vineyards. New Zealand Winegrowers (1999)
11. Molitor, D., Caffarra, A., Sinigoj, P., Pertot, I., Hoffmann, L., Junk, J.: Late frost damage risk for viticulture under future climate conditions: a case study for the Luxembourgish winegrowing region. *Aust. J. Grape Wine Res.* **20**(1), 160–168 (2014)
12. Johnson, D.E., Howell, G.S.: Factors influencing critical temperatures for spring freeze damage to developing primary shoots on concord grapevines. *Am. J. Enol. Vitic.* **32**(2), 144–149 (1981)
13. Meier, U.: Growth stages of mono-and dicotyledonous plants. In: BBCH Monograph, Federal Biological Research Centre for Agriculture and Forestry (2001)



A New Algorithm for the Imputation of Missing Functional Data

Marco Stefanucci^(✉)

Department of Economics and Finance, University of Rome, Tor Vergata,
Rome, Italy

`marco.stefanucci@uniroma2.it`

Abstract. In this work we introduce an algorithm designed for the imputation of missing functional data. The algorithm alternates between a curve imputation step and a parameter update step, resulting in an iterative procedure. The proposal is compared to well-known existing alternatives, showing outstanding results. Finally, we showcase the potential of the method using simulations and a real dataset.

Keywords: Functional Data Analysis · Missing Data

1 Introduction

Nowadays, we often deal with collections of samples represented as curves, surfaces, or other types of functional data. A challenge with such data is handling missing information, which can happen due to various reasons like instrument failure or dropout in studies.

There are two common approaches for dealing with missing data. In the first approach, called “partially observed functional data”, it is assumed that functions are observed on a regular grid shared by all units, and missing data appear as “fragments”. For instance, a function might be missing in a specific subset of the domain, but observed in the rest. The second approach is “sparse functional data”, where data for each unit are available only at specific discrete points on a common domain, often with some measurement error.

Even though both approaches aim to learn the underlying unobserved function, the methods used are usually different. In the case of partially observed data, the focus is on reconstructing curves and estimating parameters like the mean function and covariance operator. Various methods have been developed for this, such as regularized estimators and reconstruction based on eigenfunctions ([1–3]). For sparse functional data, the research is rooted in using basis expansions for modeling unevenly sampled longitudinal data ([4, 5]). The Principal Analysis through Conditional Expectation (PACE) algorithm ([6]) is a popular method in this context, aiming to predict each score of functional principal components by considering expectations conditioned on the observed part of the curve.

This work aims to develop a procedure using an iterative algorithm to estimate important population parameters, like the mean function and spectral components of the covariance operator, and simultaneously infer missing parts of functions, for both sparse and partially observed functional data. This approach could potentially be extended to handle additional covariates in structured studies.

2 A Framework for Completely Observed Data

In this section, let's assume we have a set of functional data denoted as $\mathcal{D} = \{X_i(t)\}_{i=1}^n$, where each X_i represents an independent copy of the same stochastic process $X : \mathcal{T} \rightarrow \mathbb{R}$. For simplicity, we consider the interval $\mathcal{T} = [0, 1]$. Each curve in our sample is densely observed on a common domain represented by a set of finite locations $\mathbf{t} = \{t_1, \dots, t_p\}$, essentially a discretization of $\mathcal{T}$. These observations may be affected by independent measurement error, meaning the observed data is given by $X_{ij} = X(t_{ij}) + \varepsilon_{ij}$. Our main goal is to estimate population parameters, often expressed through the Karhunen–Loeve expansion:

$$X_i(t) - \mu(t) = \sum_{k=1}^{\infty} u_{ik} V_k(t) \quad i = 1, \dots, n$$

Here, $\mu(t)$ is the mean function, $\{V_k(t)\}_{k=1}^{\infty}$ are orthonormal eigenfunctions of the covariance operator $\mathcal{R}$, and $\{u_{ik}\}_{i=1}^n$ are independent and identically distributed realizations of the scalar random variable U_k . The literature often assumes a certain level of regularity for unknown functions within the space $\mathcal{W}^k$ of continuous functions with up to k continuous derivatives. Within this framework, the mean function is estimated by solving an optimization problem, resulting in a regularized version of the cross-sectional mean known as a smoothing spline estimate.

$$\hat{\mu}(t) = \arg \min_{m \in \mathcal{W}^2} \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \{X_{ij} - m(t_j)\}^2 + \gamma \int m''(t)^2 dt. \quad (1)$$

Eigenfunctions can be estimated sequentially through the smoothing spline method too ([7, 8]). Each eigenfunction is found by solving another optimization problem, providing a regularized version of the rank-one approximation of a *deflated* matrix.

$$\{\hat{\mathbf{u}}_k, \hat{V}_k(t)\} = \arg \min_{\mathbf{u} \in \mathbb{R}^n, V \in \mathcal{W}^2} \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \{\tilde{X}_{ij} - u_i V(t_j)\}^2 + \gamma \int V''(t)^2 dt \quad (2)$$

For practical reasons, only the top K eigenfunctions associated with the largest eigenvalues are estimated. The procedure also yields predictions of the scores, from which eigenvalues can be estimated.

3 Missing Data Imputation

Regrettably, the mentioned methods encounter challenges when dealing with missing data. In fact, to estimate unknowns data is organized into a matrix $\mathbf{X}$ of size $n \times p$, where n is the sample size, and p is the length of the vector $\mathbf{t}$. This matrix is crucial for subsequent estimation processes. In case of missing data, each curve is observed on a subject-specific grid $\mathbf{t}_i = \{t_{i1}, \dots, t_{ip_i}\}$. For partially observed functional data, these points are close to each other, reflecting the discretization of the interval $\mathcal{T}_i$ where the function is observed. In contrast, in sparse settings, $\mathbf{t}_i$ consists of isolated points, and the measurement error is significant. However, data can still be stored in a $n \times p$ matrix $\mathbf{X}$, where p is the length of the vector $\mathbf{t} = \cup_i \mathbf{t}_i$, representing the discretized common domain. Since the data matrix $\mathbf{X}$ is now incomplete, computations from the previous section are not feasible. Problems (1) and (2) require a complete data matrix, and direct estimation of the covariance is hindered by missing observations. Even obtaining a smooth version of the curves from noisy data is often impossible due to limited and fragmentary observations. An immediate alternative is the “available cases analysis”, where computations are performed using only observed data, excluding missing data. However, this approach can introduce biases, reduce power, and damage the overall estimation process. Furthermore, it does not provide imputed curves.

This work explores a new approach in which the imputation of missing curves is iterated and alternated with a step where model parameters are estimated. This procedure benefits from contextual imputation and parameter estimation, improving upon two-step strategies where missing data are imputed after parameter estimation. In the multivariate context, authors have explored the use of latent structures for imputing missing values. Notably, [9] fitted a low-rank model to large data matrices with missing entries. Our approach builds on similar principles, adapting the algorithm to the functional data setting. More specifically, in the first step of the generic r -th iteration of the proposed algorithm, we fill the missing data entries using the following prediction rule based on low-rank representation of data:

$$\widehat{X}_i(t)^{(r)} = \begin{cases} X_i(t) & t \in \mathbf{t}_i \\ \widehat{\mu}(t)^{(r)} + \sum_{k=1}^K \widehat{\mathbf{u}}_{ik}^{(r)} \widehat{V}_k(t)^{(r)} & \text{otherwise} \end{cases} \quad i = 1, \dots, n \quad (3)$$

Once we have a complete matrix $\widehat{\mathbf{X}}$, in the second step we update model parameters $\widehat{\mu}(t)^{(r+1)}$, $\widehat{\mathbf{u}}_k^{(r+1)}$ and $\widehat{V}_k(t)^{(r+1)}$ iteratively for $k = 1, \dots, K$ by applying formulas (1) and (2) to the complete data matrix. We repeat this process for $r = 1, \dots, R$, where R is chosen based on some stopping rule such as a threshold on the norm of the difference between two consecutive imputation steps. The application of the algorithm varies according to the nature of the data. In the sparse setting or when data have measurement errors, the algorithm proves valuable by offering predictions for missing entries and providing a smooth representation of observed functions. Upon convergence, the final version of each

curve can be determined using the relation:

$$\widehat{X}_i(t) = \widehat{\mu}(t) + \sum_{k=1}^K \widehat{u}_{ik} \widehat{V}_k(t)$$

In the partially observed framework, where data are observed without measurement errors on a fine grid, the available part of each curve doesn't need modification or smoothing. Therefore, the final version of the working dataset is a matrix where missing entries are imputed, and observed entries remain unchanged.

4 Simulation Study and Real Data Analysis

In this section, we assess the practical performance of the proposed method in both partially observed and sparse data settings. Our focus is on evaluating the algorithm's ability to recover information from a fully observed sample. Data generation follows these steps: for each scenario, we create the mean function $\mu(t)$ and K orthogonal functions $V_1(t), \dots, V_K(t)$ using a B-spline basis with L elements, and coefficients from a standard normal distribution. Completely observed functions are then derived from the low-rank model:

$$X_i(t) = \mu(t) + \sum_{k=1}^K u_{ik} V_k(t) + \varepsilon_i \quad i = 1, \dots, n$$

Here, $u_{ik} \sim N(0, \lambda_k)$, and $\varepsilon_i \sim N(0, \sigma^2)$. These data remain constant across simulation repetitions. Two scenarios are considered. The first is a partially observed scenario without measurement error, $K = 6$ active eigenfunctions, and eigenvalues $\lambda = (1.23, 1.04, 0.71, 0.47, 0.27, 0.08)$. In the sparse scenario, we use a signal-to-noise ratio equal to 10, $K = 3$, and $\lambda = (2.79, 1.11, 1.05)$. The missing data mechanism that we consider in the first scenario is the following: for each curve in the dataset, a random segment of Lebesgue measure 0.3 is missing. These segments are independent across units and simulation repetitions and also, independent from data. In the sparse scenario, we assume each curve is observed at 10 random points in the domain. To evaluate our algorithm, we calculate the error for various parameters, comparing the estimate obtained by the algorithm using missing data to the one obtained with the fully observed sample. The parameters include the mean function $\mu(t)$, the eigenfunctions $V_1(t), \dots, V_K(t)$, and the eigenvalues $\lambda_1, \dots, \lambda_K$. Errors are computed as the Integrated Mean Squared Error (MISE) for functions and Mean Squared Error (MSE) for scalars. We conduct a comparison of our results with those achievable using PACE. Additionally, we calculate the average error in imputing a curve, utilizing the Integrated Mean Squared Error (MISE), and compare our outcomes with both PACE and the Kneip and Liebl approach. Results are presented in Tables 1, 2

and 3. We conclude that the advantage of the proposed method is clear in both estimation and imputation, for all considered scenarios. We did not experience lack of convergence of the method in our simulations.

Table 1. Average Integrated Mean Squared Error (first row) and Average Mean Squared Error (second row, $\times 10^3$), with standard deviation in parenthesis, attained by our method and PACE for several parameters, scenario 1.

	μ	V_1	V_2	V_3	V_4	V_5	V_6
NEW	0.47 (0.22)	3.47 (4.01)	3.95 (4.12)	2.28 (2.11)	3.06 (2.24)	2.98 (1.79)	1.96 (0.53)
PACE	2.39 (0.93)	38.19 (48.62)	40.91 (47.63)	10.15 (8.20)	15.31 (18.11)	14.53 (15.95)	24.74 (17.57)
		λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
NEW		0.34 (0.48)	4.41 (4.68)	2.15 (2.55)	1.26 (1.69)	6.87 (2.98)	0.77 (0.30)
PACE		13.96 (19.27)	17.47 (19.50)	8.36 (9.43)	5.08 (5.74)	3.69 (4.25)	0.87 (0.89)

Table 2. Average Integrated Mean Squared Error and Average Mean Squared Error ($\times 10^3$), with standard deviation in parenthesis, attained by our method and PACE for several parameters, scenario 2.

	μ	V_1	V_2	V_3	λ_1	λ_2	λ_3
NEW	0.09 (0.05)	0.03 (0.02)	0.38 (0.67)	0.40 (0.67)	2.33 (2.88)	1.06 (1.26)	1.01 (1.22)
PACE	2.33 (1.26)	1.27 (1.10)	22.47 (29.99)	23.19 (30.03)	56.55 (68.58)	46.12 (66.69)	46.17 (53.24)

Table 3. Average Integrated Mean Squared Error, with standard deviation in parenthesis, attained by all methods in imputation.

	NEW	PACE	KL
Scenario 1	21.66 (5.29)	54.26 (27.3)	38.89 (54.61)
Scenario 2	4.23 (2.11)	37.89 (49.88)	28.16 (12.07)

We further explore the potential of the method by examining a meat absorbance spectrum dataset named Tecator, a commonly used benchmark in functional data analysis. In its original form, this dataset does not contain missing data. However, in this experiment, we artificially introduce missingness and attempt to reconstruct the deleted portions of curves. Unlike the simulations, we lack control over population parameters such as the number of relevant eigenfunctions and the rate of decay of the eigenvalues. We found the model with $K = 4$ to be the one that minimizes the imputation error. The overall outcome is highly satisfying, as demonstrated in Fig. 1, where some imputed curves are presented.

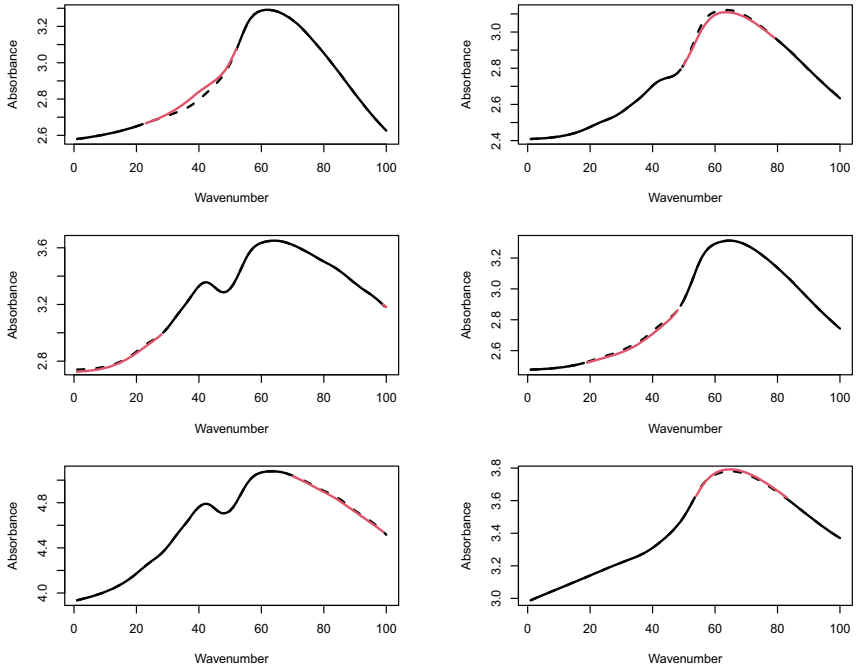


Fig. 1. Randomly selected imputed curves in the Tecator dataset.

5 Conclusions

In this study, we introduced an innovative iterative algorithm with the dual purpose of inferring missing parts of curves and estimating population parameters. Drawing inspiration from the matrix completion method outlined in [9], we extended the procedure to handle functional observations. The proposal stands out in comparison to other methods in various aspects, suggesting potential for further research and broader applicability.

Acknowledgements. This research was partially supported by the MUR-PRIN 2022 project CARONTE (Prot. 2022KBTEBN), funded by the European Union - Next Generation EU.

References

1. Kraus, D.: Components and completion of partially observed functional data. *J. R. Stat. Soc. Ser. B Stat Methodol.* **77**, 777–801 (2015)
2. Kraus, D., Stefanucci, M.: Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika* **106**, 161–180 (2018)
3. Kneip, A., Liebl, D.: On the optimal reconstruction of partially observed functional data. *Ann. Stat.* **48**, 1692–1717 (2020)

4. James, G., Hastie, T., Sugar, C.: Principal component models for sparse functional data. *Biometrika* **87**, 587–602 (2000)
5. James, G.M., Sugar, C.A.: Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.* **98**, 397–408 (2003)
6. Yao, F., Müller, H.-G., Wang, J.-L.: Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.* **100**, 577–590 (2005)
7. Huang, J.Z., Shen, H., Buja, A.: Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.* **2**, 678–695 (2008)
8. Lila, E., Aston, J.A.D., Sangalli, L.M.: Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.* **10**, 1854–1879 (2016)
9. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11**, 2287–2322 (2010)



Exploring Disease Prevalence in Italy: A Web Application Utilizing Sample-Based Predictions from a Bayesian Spatio-Temporal Model with External Validation

Mattia Stival^(✉), Lorenzo Schiavon, Gaia Bertarelli, and Stefano Campostrini

Department of Economics, Ca' Foscari University of Venice, Venice, Italy
`mattia.stival@unive.it`

Abstract. Disease prevalence data is an invaluable resource in population monitoring due to its easy representation and immediate interpretability. However, accessing such information can often be challenging. Although, sampling methods provide an efficient alternative for inferring population prevalence and collecting correlated risk factor information, challenges arise due to potential under representation, necessitating robust modeling approaches. Hierarchical Bayesian models may offer flexible solution to integrate prior knowledge and considering individual and environmental factors in disease prevalence estimation. However, interpreting complex statistical models poses challenges for non-experts. To address this, a user-friendly web interface leveraging Shiny is proposed, facilitating comparison and validation of disease prevalence estimates. This infrastructure aims to enhance accessibility and reliability, with potential for broader applicability and future improvements.

Keywords: health surveillance systems · disease prevalence · PASSI dataset

1 Introduction

Despite their great relevance in population monitoring, disease prevalence data are often difficult to access. Sample data may represent an efficient alternative for inferring population prevalence and collecting further information possibly correlated. In this regard, various nations, including Italy, have developed notable methods for monitoring population health, which involve conducting frequent cross-sectional surveys at brief time intervals. These surveys, commonly known as BRFSS (Behavioral and Risk Factors Surveillance system), collect data on the well-being of individual participants, encompassing details about their behaviors, risk factors, and pertinent socio-demographic information [1]. In Italy, PASSI (*Progressi delle Aziende Sanitarie per la Salute in Italia*) is a surveillance system that, since 2008, collects sample data on the behavioral and risk factors of

the Italian population in the age span of 18 to 69 years at the level of Local Health Units (ASL) [2]. Given its considerable importance, various authors have analyzed the BRFSS datasets [1] [5]. For Italy, recent works [3][4] have utilized the PASSI dataset to model the age-morbidity relationship and its association with socio-demographic risk factors. However, the potential of the PASSI system extends beyond that, collecting valuable information at a local level that can be utilized to investigate the heterogeneity within the country. However, operating at finer levels with sample data could present challenges, as certain segments of the population may be underrepresented within the sample. Therefore, appropriate modeling is necessary to effectively leverage the available information across the entire dataset and to enhance the estimation process through the incorporation of prior knowledge. A hierarchical Bayesian approach may offer a solution to address this challenge, enabling the integration of additional prior information to refine morbidity estimates. In these concerns, [10] propose a Bayesian multivariate spatio-temporal logistic model for chronic disease diagnoses, that can be applied not only at the national level but also at finer resolutions such as ASLs. The model incorporates the influence of both individual and environmental risk factors on the marginal probability of developing specific chronic diseases. Additionally, it includes a latent comorbidity indicator that accounts for the correlation between diseases due to an individual's propensity for comorbidity.

While survey methods offer lower costs and quicker access to information compared to alternative monitoring methods, estimates derived from survey samples, often obtained through complex statistical models, may prove difficult for non-statistical experts to access. Furthermore, despite typically undergoing validation using internal methods, there's no assurance that sample-based estimates and predictions accurately reflect more reliable population data. This preliminary work aims to achieve two main objectives: (i) develop and present a user-friendly web interface using Shiny [8] for comparing disease prevalence among small population subgroups, building upon the work by [10]; (ii) establish a framework for validating predictions using aggregated population data from external sources. The proposed infrastructure aims to improve the reliability and accessibility of morbidity predictions derived from a sophisticated statistical model [10], catering to audiences with diverse backgrounds. This enables simplified yet accurate and reliable analyses. Additionally, the validation procedure may serve as a fair assessment of accuracy, possibly facilitating comparisons between predictions from different models to identify their strengths and weaknesses.

2 Overview of the Model and External Validation

Following [10], for a generic respondent i ($i = 1, \dots, n_s$), we consider the n_d -variate binary response $\mathbf{y}_i$ indicating the presence or absence of n_d possible diseases, with $y_{ij} = 1$ if the disease j has been diagnosed to the subject i . For each respondent i , an n_p -variate vector $\mathbf{x}_i$ of covariates, representing individual

non-modifiable risk factors, is also available, with h -th element x_{ih} representing the value of covariate h . Let $\text{Ber}(\pi)$ denote a Bernoulli random variable with mean π . Then, the logistic model

$$y_{ij} \sim \text{Ber}(\pi_{ij}), \quad \pi_{ij} = \text{logit}^{-1}(\eta_{ij}),$$

has been defined. The continuous latent random vector $\boldsymbol{\eta}_i = [\eta_{i1}, \dots, \eta_{in_d}]^\top$ is defined as

$$\boldsymbol{\eta}_i = B(l_i, c_i)\mathbf{x}_i + \boldsymbol{\gamma}\epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad (1)$$

with mean $E(\boldsymbol{\eta}_i) = B(l_i, c_i)\mathbf{x}_i$ and residual term $\boldsymbol{\gamma}\epsilon_i$. In equation (1), $B(l_i, c_i)$ is a $n_d \times n_p$ matrix of coefficients depending on the location l_i and the cohort c_i to which the respondent i belongs to. Using a state space formulation, spatio-temporal dependence is included in the model by considering a dynamic-linear evolution for the terms in $B(l, c)$ and exploiting exogenous information available for the locations. The term ϵ_i is a scalar morbidity index accounting for propensity to co-morbidity of the respondent, scaled for each disease according to the parameter vector $\boldsymbol{\gamma}$. For reasons of space we do not report the entire specification of prior distributions. Please refer to [10] for complete details.

Given the model, the predictive distribution for a new subject l based on the set of sample data $\mathcal{D}$ is

$$p(\mathbf{y}_l | \mathcal{D}) = \int \prod_{j=1}^{n_d} \pi_{lj}^{y_{lj}} (1 - \pi_{lj})^{1-y_{lj}} p(\boldsymbol{\pi}_l | \mathcal{D}) d\boldsymbol{\pi}_l,$$

with $\mathbf{y}_l = (y_{l1}, \dots, y_{ln_d}) \in \{0, 1\}^{n_d}$. When a sample of $p(\boldsymbol{\pi}_l | \mathcal{D})$ is available (with $\boldsymbol{\gamma}$ and ϵ_l marginalized out), the individual predictive distribution $p(\mathbf{y}_l | \mathcal{D})$ (or samples from it) can be recovered via Monte Carlo integration. Further, by marginalization, the distribution $p(y_{lj} = 1 | \mathcal{D}) = \sum_{\mathbf{y}_l \in \{0,1\}^{n_d}} p(\mathbf{y}_l | \mathcal{D}) \mathbb{1}(y_{lj} = 1)$ is also available.

For a generic population subgroup $\mathcal{C}_c$, the predictive distribution of prevalence in probability scale for disease j can be computed as

$$f_{jc} = \sum_{l \in \mathcal{C}_c} w_{lc} p(y_{lj} = 1 | \mathcal{D}),$$

with $w_{lc} \in (0, 1)$ and $\sum_{l \in \mathcal{C}_c} w_{lc} = 1$ representing the probability of a subject in the population having the characteristics of unit l , given its belonging to class $\mathcal{C}_c$. In this work, we assume random sampling, so that $w_{lc} = 1/F_c$, with F_c representing the number of units in the sample belonging to the population subgroup $\mathcal{C}_c$. More accurate specifications are left for future work.

2.1 Routine for External Validation Using Proper Scoring Rules

While the model described by [10] is estimated and validated using PASSI dataset, in this work we propose to evaluate model's predictions considering

external sources of data. Given the availability of external population prevalence data for a group of ASLs, we compute the observed population prevalence f_{jc}^{obs} of each disease for each year within the range of 2010 to 2020, stratifying the population by sex. Subsequently, we assess the model predictions by comparing the credible intervals obtained from f_{jc} with f_{jc}^{obs} , where the population subgroups $\mathcal{C}_c$ are defined by sex, year of survey, and ASL. Since the entire distribution of f_{jc} is available from the model, we opt to assess the predictions using proper scoring rules [6]. These rules provide summary measures for the evaluation of probabilistic forecasts by assigning a numerical score based on the predictive distribution and the observed outcome, enabling the ranking of competing predictions and the identification of deficiencies. We propose evaluating the predictions based on the Continuous Ranked Probability Score (CRPS) and interval score (IS), computed from distributions derived from Monte Carlo samples [7]. CRPS rewards predictive distributions that place mass close to the realized outcome (sensitivity to distance [7]), while IS rewards accurate interval predictions with narrow bands. Lower scores in both cases indicate better predictions. By aggregating these synthetic indices by ASL, one may rank the ASLs to compare their specific “external reliability”.

3 Overview of the Web-App

The web application infrastructure is developed using Shiny [8], an R package that simplifies the creation of interactive web applications directly within the R environment [9]. The structure of the website is illustrated in Fig. 1. The website header (A) provides information about the samples used (birth cohorts considered), along with essential bibliographic details of the work. The navigator on the right (B) enables users to select the subgroup of the population of interest, from which predictive distributions are computed. Filtering options include regional and ASL affiliation, sex, educational status, self-declared economic level, and smoking status (smoker and former smoker vs non-smoker). To ensure efficient functionality, a thinned sample of size 375 from the predictive distributions was pre-computed for each individual in the PASSI dataset considered by [10], and linked to their respective vectors of anonymized individual information. Once the population subgroup is selected, users can simply press the ‘Display curves’ button to obtain the desired prevalence curves. For each disease, these curves represent the evolution of prevalence over years. Using the stratification option in (D), they can be displayed at the aggregate level or stratified by a single individual characteristic. The curves are then displayed in the main body of the website (C). Limiting users to only one stratification option allows for comparisons of interest with sufficiently large sample guarantees. When these guarantees are not met, an information banner, shown below in Fig. 1, is displayed to the user in (E). The information banner is shown whenever samples with a size smaller than 30 are considered. Curves are not displayed when stratification results in population subgroups without sampled units. This last aspect has to be considered as the main limitation of the proposed web-app and it is a direct consequence

of how the mixture weights w_{ic} were set. Addressing this issue would enable the full benefits of Bayesian statistical modeling combined with sampling survey techniques, providing estimates and relative uncertainty quantification even for very small population subgroups. This aspect, that requires methodological attention, is beyond the scope of this preliminary work.



Fig. 1. On the top, the web-app with rectangular boxes highlighting its relevant sections. On the bottom, the informative banner displayed when some points of the curves are obtained using a sample size with less than 30 respondents.

4 Discussions and Future Works

While essential for population monitoring, population disease prevalence data are often difficult to obtain. Conversely, sample-based prevalence, while easier to recover, may suffer from underrepresentation issues for small population segments. Bayesian hierarchical models allow for the utilization of survey data potential, enabling the estimation of morbidity curves and uncertainty quantification, even with limited data, particularly when dealing with inherently small population groups.

This work presents the procedure to derive annual prevalence curves for any population subgroup of interest, based on morbidity predictions from Bayesian spatio-temporal models. These estimates are derived by combining individual predictive distributions. Additionally, we propose a validation methodology leveraging external prevalence data when available. To enhance accessibility, we have developed a user-friendly web application. This tool facilitates data filtering for any population subgroup based on individual variables such as sex, educational level, economic status, smoking status, and location information. Users

can also specify up to one stratification variable for comparisons. At this preliminary stage, random sampling is assumed, utilizing equal mixture weights for prevalence computation. While reasonable for PASSI data, this approach has the drawback of excluding population subgroups that are too small to be sampled, thus only partially leveraging the advantages of Bayesian spatio-temporal models.

We plan to address this limitation in future work by exploiting the potential of Bayesian models for frequency estimation using survey data—see e.g. [11].

Funding Information. This long abstract was developed within the project funded by Next Generation EU - Age-It - Ageing well in an ageing society project (PE00000015), National Recovery and Resilience Plan (NRRP) - PE8 - Mission 4, C2, Intervention 1.3. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

1. Assaf, S., Campostrini, S., Xu, F., Crawford, C.G.: Analysing behavioural risk factor surveillance data by using spatially and temporally varying coefficient models. *J. Roy. Stat. Soc. Ser. A: Stat. Soc.* **179**, 153–175 (2015)
2. Baldissera, S., et al.: Peer reviewed: features and initial assessment of the Italian behavioral risk factor surveillance system (passi), 2007–2008. *Prevent. Chronic Dis.* **8**, A24 (2011)
3. Pastore, A., Tonellato, S., Aliverti, E., Campostrini, S.: When does morbidity start? an analysis of changes in morbidity between 2013 and 2019 in Italy. *Stat. Methods Appl.* **32**, 577–591 (2023)
4. Andreella, A., Monasta, L., Campostrini, S.: A novel comorbidity index in Italy based on diseases detected by the surveillance system passi and the global burden of diseases disability weights. *Popul. Health Metrics* **21**, 18 (2023)
5. Zheng, X., Zhang, X., Jorge, C., Aye, D.: Model-based community health surveillance via multilevel small area estimation using state behavioral risk factor surveillance system (BRFSS): a case study in connecticut. *Ann. Epidemiol.* **78**, 74–80 (2023)
6. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**(477), 359–378 (2007)
7. Kruger, F., Lerch, S., Thorarinsdottir, T., Gneiting, T.: Predictive inference based on Markov chain Monte Carlo output. *Int. Stat. Rev.* **89**(2), 274–301 (2021)
8. Chang, W., et al.: shiny: Web Application Framework for R. R package version 1.7.3 (2022)
9. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022)
10. Stival M., Schiavon L., Campostrini S.: A Bayesian approach to uncover spatio-temporal determinants of heterogeneity in repeated cross-sectional health surveys. [Arxiv:2402.19162](https://arxiv.org/abs/2402.19162) (2024)
11. Lui, K.-J., Cumberland, W.G.: A Bayesian approach to small domain estimation. *J. Off. Stat.* **5**(2), 143–156 (1989)



Dependent Infinite Latent Feature Models for Tree Occurrence Data

Federica Stolf^(✉)

Department of Statistical Sciences, University of Padova, Padova, Italy

`federica.stolf@phd.unipd.it`

Abstract. Recent advancements in species sampling methods and bioinformatic tools have led to the collection of many species occurrence datasets, motivating the development of statistical tools to extract valuable insights and enhance our understanding of biodiversity. There is a rich literature in ecology on so-called joint species distribution models (JSDMs), which usually take the form of multivariate probit latent factor regression models. However, such models cannot deal with the fact that we regularly discover many new species as the sampling is being conducted. In this paper, we analyze tropical tree occurrence data leveraging an infinite latent features model. This approach does not require specifying species identity in advance and allows a growing number of binary outcomes.

Keywords: Multivariate binary response · Bayesian factor model · Indian buffet process · Ecology

1 Introduction

Species sampling data consist of a $n \times p$ matrix Y of binary indicator of species occurrences, where $y_{ij} = 1$ if species j was found in sample i and 0 otherwise, for $i = 1, \dots, n$ and $j = 1, \dots, p$. The main interests are to infer the dependence structure across the different species and to estimate marginal species occurrence probabilities. A popular tool in ecology for this problem is Joint Species Distribution Models (JSDMs) [1], which model the data in a multivariate binary regression framework. In general, they achieve good predictive performance, but they focus only on a prespecified finite number of species. This could be a major limitation since the number of species sampled is unknown a priori and often very large. Consequently, the analysis focuses only on common species, while the number of rare species is huge in practice and they have a critical impact in the understanding of biodiversity.

Indian Buffet Processes (IBPs) [2] for latent feature modeling in Bayesian nonparametrics are a natural tool to allow a growing number of binary outcomes. The IBP is a distribution for sampling binary matrices with any finite number of rows (sample) and an unbounded number of columns (features). In our motivating application the latent features correspond to which species are discovered

in a sample. However, the IBP makes unrealistic assumptions of independence in species occurrence. To account for within-sample dependence we rely on the new class of dependent infinite latent feature models, the multivariate probit Indian buffet process (MVP-IBP), recently proposed by [3]. This method, differently from the IBP, models the dependence among species through a correlation matrix as in a multivariate probit fashion [4]. The MVP-IBP provides a general framework that bridges between multivariate probit models and IBPs. In this paper, we focus on a special case of this approach, the hierarchical MVP-IBP factor model, where the dependence among species is modeled through a factor hierarchical structure. This formulation for the correlation matrix allows at the same time to effectively reduce the dimensionality and to borrow information across species, to obtain accurate estimation also for rare species.

2 Model Specification

Let Y be a binary feature matrix, with n rows and an unbounded number of columns, where $y_{ij} = 1$ indicates the possession of feature j by object i . The IBP [2] can be derived as the infinite limit of a beta-Bernoulli model. Consider the case of finite feature models with p known number of features, such that Y is a $n \times p$ matrix. The IBP of parameter $\alpha \in \mathbb{R}^+$ is derived by taking the $p \rightarrow \infty$ limit of the conditional distributions

$$\pi_j \sim \text{Beta}(\alpha/p, 1), \quad y_{ij} \mid \pi_j \sim \text{Bernoulli}(\pi_j), \quad (1)$$

where the π_j are generated independently and each y_{ij} is independent of all other assignments conditioned on π_j . In the Indian buffet metaphor, rows of Y correspond to customers and columns to dishes and the binary matrix Y is sampled as a sequential process. In this process, customers will tend to choose popular dishes; hence as n grows Y gets sparser.

The MVP-IBP [3] model corresponds to a probit-Bernoulli process, having a construction similar to the IBP. As in the multivariate probit setting [4], the MVP-IBP is defined through a latent Gaussian construction. Let $z_i = (z_{i1}, \dots, z_{ip}) \in \mathbb{R}^p$ denote a p -variate latent continuous variable for $i = 1, \dots, n$ and let y_{ij} be 1 or 0 according to the sign of z_{ij} . The hierarchical factor MVP-IBP model is defined as

$$y_{ij} = \mathbb{1}(z_{ij} > 0), \quad z_{ij} = \beta_j + \lambda_j^T \eta_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 1), \quad (2)$$

where $\beta = (\beta_1, \dots, \beta_p)^T$ is a p -dimensional vector of coefficients controlling the marginal dish selection, $\lambda_j^T \eta_i$ is a factor analytic term inducing dependence, and $\mathbb{1}(x)$ is the indicator function of x . Here, $(\eta_{i1}, \dots, \eta_{ik})^T \sim N_K(0, I)$ is a $k \times 1$ vector of latent variables and $\lambda_j^T = (\lambda_{j1}, \dots, \lambda_{jk})$ denotes the j th row of Λ a $p \times k$ matrix of factor loadings ($k \ll p$). The factor loadings are interpreted as species latent attributes, while the latent variable as characteristics of the i th sample. For example, η_i may measure habitat attributes at that sampling location and calendar time, while λ_j measures the impact those habitat attributes have on

the occurrence of species j . The dependence among the components is induced by marginalizing over the distribution of the factors, so marginally $z_i \mid \beta, \Lambda \sim N(\beta, \Omega)$ where $\Omega = \Lambda\Lambda^\top + I$.

To obtain a parametrization in terms of the correlation matrix, as standard practice for multivariate probit model [4], we follow [3] and scale the latent variable z_i . Thus, inference takes place on the scaled coefficients

$$\tilde{\beta}_j = \frac{\beta_j}{\sqrt{1 + \lambda_j^\top \lambda_j}}, \quad (3)$$

for which we take a normal prior $\tilde{\beta}_j \sim N(\mu_p, \tau_p^2)$ with the default MVP-IBP value for the parameters μ_p and τ_p^2 [3]. The formulation for μ_p depends on a parameter α , which plays the same role as the IBP parameter. For a comprehensive study of the theoretical properties of the MVP-IBP model when $p \rightarrow \infty$ refer to [3].

We assume a hierarchical structure for the factor loadings to account for rare species and borrow information among species. Specifically, we let $\lambda_j^\top \sim N_k(\nu, \Omega)$ and we take a normal–Inverse–Wishart prior for (ν, Ω) . The number of factors can be chosen to optimize some goodness-of-fit criteria, such as the Watanabe–Akaike information criterion. To complete a Bayesian specification of the model, we take $\alpha \sim Ga(a_\alpha, b_\alpha)$.

3 Tropical Forest Trees Application

We applied the hierarchical factor MVP-IBP model to tree occurrence data. These data come from a community ecology survey conducted in the tropical moist forest of Barro Colorado Island (BCI) in Gatun Lake in central Panama [5] and they are public available in the R package `vegan`. The data consists of 225 different species of trees observed in $n = 50$ sampling areas.

Posterior inference for the hierarchical factor MVP-IBP model truncated at p terms for the features proceeds via a Gibbs sampler. This sampler relies on a data augmentation which exploits the latent variable construction of the probit model. The Gibbs sampler is combined with a Metropolis-Hastings approach for updating the factor loadings and the α parameter. Specifically, to update the factor loadings λ_j we use a Metropolis adjusted Langevin algorithm (MALA) [6].

To assess the performance of the model in estimating marginal tree occurrence probabilities, we draw 50 samples from the predictive distributions of Y and average across the predictive samples. Figure 1 shows the predictive point-wise posterior mean for tree occurrences obtained with the hierarchical factor MVP-IBP model. For graphical clarity, the plot includes only 50 randomly chosen species. The model’s recovery of the binary matrix agrees with the observed data. Additionally, in community ecology is of interest to understand how species interact. For instance, a positive correlation between two species indicates favorable conditions for both within the habitat. To this aim, we can estimate the posterior mean for the correlation matrix averaging across the Markov chain

Monte Carlo (MCMC) samples. The results relative to the twenty more common tree species are shown in Fig. 2, revealing several species with positive correlation. The positive association among trees plays a key role for ecologists in biodiversity mapping.

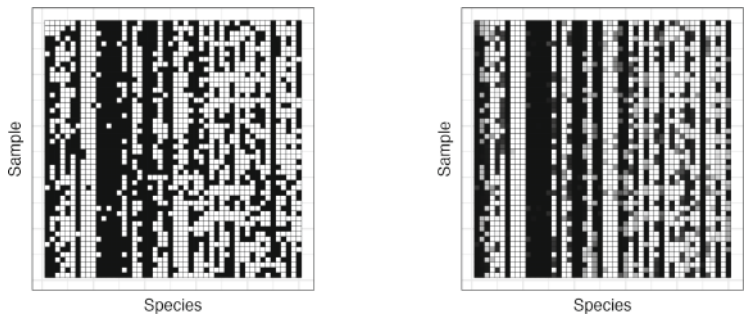


Fig. 1. Tropical tree occurrences matrix (left panel) and posterior predictive mean obtained with the MVP-IBP factor model (right panel) of 50 tree species.

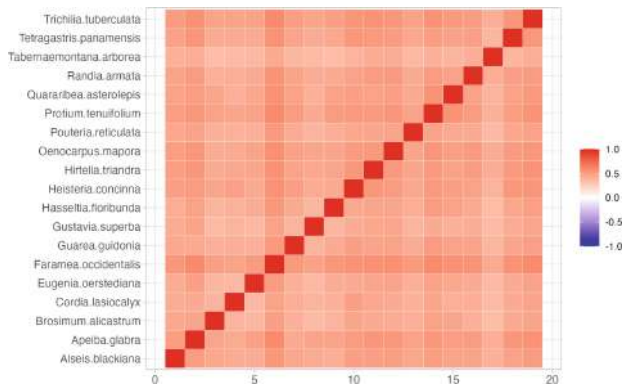


Fig. 2. Posterior mean of correlation coefficients estimated by the hierarchical factor MVP-IBP model for the twenty most common tree species.

References

1. Ovaskainen, O., Abrego, N., Halme, P., Dunson, D.B.: Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* **7**(5), 549–555 (2016)

2. Griffiths, T., Ghahramani, Z.: The Indian buffet process: an introduction and review. *J. Mach. Learn. Res.* **12**(32), 1185–1224 (2011)
3. Stolf, F., Dunson, D. B.: Allowing growing dimensional binary outcomes via the multivariate probit Indian buffet process. arXiv preprint [arXiv:2402.13384](https://arxiv.org/abs/2402.13384) (2024)
4. Chib, S., Greenberg, E.: Analysis of multivariate probit models. *Biometrika* **85**(2), 347–361 (1998)
5. Condit, R., et al.: Beta-diversity in tropical forest trees. *Science* **295**(5555), 666–69 (2002)
6. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Stat. Soc. Ser. B: Stat. Methodologl.* **73**(2), 123–214 (2011)



War and the Military Service in Post-Communist Bloc: Which Impact on Marriage and Divorce?

Sara Tafuro^(✉)

Università degli Studi di Firenze, Firenze, Italy
sara.tafuro@unifi.it

Abstract. Armed conflicts are a disruptive life event and, as such, they have the potential to influence relationship patterns. Nevertheless, the influence of wars on marriage and divorce is understudied. The limited research on this topic focuses on the United States, generally highlighting an adverse effect of armed conflicts on marital bonds. In this study, we focus on former Soviet countries employing the Life in Transition Survey (LITS), a unique longitudinal dataset providing data for individuals in Eastern Europe and Central Asia from 1989 to recent times. Hence, the period during and after the collapse of communist regimes and its numerous ensuing wars. This is a relevant case study since its socio-demographic and cultural context is widely different than the US one. Indeed, conservative family attitudes are believed to have resurfaced after the collapse of the Soviet regimes: have the post-Soviet wars directly fostered these attitudes (hence increasing marriages and decreasing divorces) or have they rather had the same disruptive effect on couples observed by previous literature in other geographical contexts?

We test through linear probability models how serving in the army and/or living in a country at war is associated with the probability of getting married or divorcing on the year before, that same year, a year later, two and three years later. Results paint a complex picture: although participating in the army seems, in general, to be positively associated with marrying and staying married, it is less the case when individuals serve in times of conflict than in times of peace.

Keywords: uncertainty · war · divorce

1 Introduction

Armed conflicts are inevitably accompanied by a pervasive sense of uncertainty and affliction which cannot but affect individuals' life course and intimate life. The scientific evidence around the impact on marriage and divorce rates of such a stressful and all-consuming life experience is, nonetheless, limited.

Hypothetically, war could affect relationships in two main contrasting ways. On one hand, the emotional stress caused by war could negatively impact on individuals' mental health, first (Cesur et al., 2013), and on the quality of their marital satisfaction later (Cook et al., 2004), possibly resulting in less marriages and more breakups. By contrast, a less explored hypothesis claims that, in order to face the violence and socio-political

uncertainty in their environment, individuals may stick together to pool their individual resources and soothe each other's emotional distress.

Most of the existing literature exploring the impact of war on marriage and divorce suggests that it either disrupts couples' stability (Angrist & Johnson, 2000; Janmardy, 2011; Negrusa et al., 2014; Pavalko & Elder, 1990), or has no effect at all (Cook et al., 2004; Lester, 1993). A couple of studies find that increasing the duration of deployment among military families is associated with a higher divorce risk (Angrist & Johnson, 2000; Negrusa et al., 2014). Nonetheless, the socio-economic, demographic and cultural circumstances in which a specific armed conflict occurs seems to play a major role in determining whether marital bonds will be strengthened, weakened or unimpacted afterwards. For example, it has been highlighted that WWII raised the divorce rates in the US whereas other episodes (e.g. the Vietnamese and Korean wars) did not: the most likely explanation lies in the fact that the increase in female labor force participation observed in WWII (in consequence of the massive deployment of troops and the ensuing want of workforce) determined a substantial financial empowerment for women and eventually decreased their interest in staying married (Anderson & Little, 1999).

To fully grasp how wars affect union formation and dissolution, it is crucial to investigate various historical and geographical context. The existing literature, however, rarely covers countries other than the United States, mostly because of data availability. What we propose to do in this paper is to exploit rare information on a context that is in stark contrast to the ones most often explored by the literature on war and union formation and dissolution. The Life in Transition Survey, indeed, offers longitudinal information on individuals and their life events in 29 countries belonging to the former communist bloc. The survey covers a very interesting time frame, ranging from 1989 (i.e. right before the regimes collapsed) up to the mid-2006 (i.e. when the data was collected): hence, almost two decades of strong political instability which saw the outbreak of numerous wars (e.g. the Yugoslav Wars, the Albanian Civil War, etc.).

The socio-demographic and cultural context of the former communist bloc is also unique. In fact, most Soviet countries adopted throughout their history a liberal attitude towards family law and, in particular, divorce – the USSR, for example, was the first country in the world to introduce the “no-fault” legal ground for divorce (Zelder, 1993) –. The divorce rates in several of these countries, previously among the highest in the world, decreased drastically after the dissolution of communist regimes. It is believed that behind such trend is a resurgence of religious values (which had been repressed during the communist era) and, consequently, of conservative family attitudes (Resetar & Berdica, 2013).

To our knowledge, no study has tested yet the role played specifically by the post-Soviet wars on union formation and dissolution trends. On one hand, it is possible that these events directly fostered familistic attitudes: driven by nationalist claims, these wars had often a religious component (see, for example, the Yugoslav wars). On the other hand, the violence and angst that characterized these conflicts might have threatened marital harmony, eventually increasing breakups (as other warfare episodes did in other geographical contexts). In this second hypothesis, thus, the post-Soviet wars might have had the direct effect of decreasing marriages and increasing divorces, even though the general trends in that period went in the opposite direction.

Hence, two conflicting hypotheses arise. The LITS dataset allows us to test them both and, interestingly, from different angles. In fact, we are able to study how individuals' likelihood of marrying and divorcing were influenced by both a direct and an indirect experience of war. Experiencing an armed conflict can, indeed, affect couples' lives simply as the event occurs in their environment (e.g. in their country, but with no member of the couple directly involved in the conflict) or, arguably even more, as one couple's member actively participates in the war (i.e. by serving in the army). In order to test the environmental effect of war, we construct a variable that identifies all observations referring to a country and a year of war. Concerning the effect of a personal involvement in a conflict, this is where the LITS dataset provides first-time information. Among the retrospective data collected yearly for its respondents, the survey reports if and when they underwent major family events (such as marrying and divorcing), but also if and when they served in the military. With these two variables combined, we are able to study thoroughly the impact of conflicts on couples' formation and dissolution at different points in time, i.e. at various years before, during and after a war.

2 Methodology

In this study, we use the Life in Transition Survey (LITS), a unique longitudinal dataset providing information for 1000 families in each of the 29 former communist countries included in the survey. The survey provides retrospective information from 1989 to 2006. This wave of the survey asks one individual in each household, for each year, about specific family events, economic activities and more: namely, if they served in the army, married or divorced, etc. This information is available for 29000 individuals across 18 years.

We test through panel linear probability models how serving in the army and/or living in a country at war (at time t) influenced individuals' probability of getting married on the year before ($t-1$), that same year (t), a year later ($t+1$), two and three years later ($t+2$ and $t+3$). While the information of serving in the army is provided in the dataset, the variable for living in country at war is a dummy we built by reviewing the war history of each country included in the dataset. An interaction of these two variables (i.e. serving in the army and living in a country at war) is also included in the regressions in order to better identify the observations of those who actively participated in the conflict (i.e. served in the military during a time of war and not simply as trainees). The same models are also run using divorce as the outcome variable.

The regressions include also various control variables, notably, time-varying indicators describing the previous marital status of respondents, their age, whether they had a kid and a measure of their economic conditions that we developed through principal component analysis. For the PCA, we used information on whether respondents were unemployed and/or in vulnerable employment and whether they asked for financial assistance from their relatives in the baseline year.

3 Results and Discussion

The regressions' main results are shown in the tables below. All models are controlled for the variables listed above.

Concerning the regressions on marriage (Table 1), let us comment first on the results of the regressions testing the effect of serving in the army and living in a country at war, before we add the interaction variable in the next step. In general, living in a country at war is associated to a significantly lower probability of marrying, during and after the conflict, compared to living in times of peace (Models 2a, 3a, 4a and 5a). The same statistically significant effect is observed also for the year preceding the war (Model 1a): individuals would hence marry less in a year preceding an armed conflict in their country, which could be due to the stress of foreseeing the upcoming crisis negatively affecting the quality of their relationships. As for serving in the army, Model 1a shows that individuals who served had a significantly lower probability of marrying on the year preceding the conscription. Later, though, the effect becomes positive: in particular, individuals who served have a systematically higher probability of marrying than those who did not, and that is one, two and three years after participating in the military (Models 3a, 4a and 5a, respectively).

From these regressions, therefore, experiencing war in one's environment seems to decrease the chances of marrying, while being in the army increases them. However, it is important to note that, in our sample, many of the individuals who served in the military did so as trainees in times of peace. While in the army, they might have received an influence of conservative values (which eventually showed in their higher chances of marrying), but without the trauma of actively participating in a war. We now add, thus, an interaction variable which identifies those who served in the army during an armed conflict and, possibly, had a first-hand experience of war.

In Model 4b, the interaction variable shows that individuals who served in times of war had a higher probability of marrying two years later than those who did not serve. At the same time, such probability is also lower than for those who served in times of peace (the coefficients are 0.0169 and 0.0555, respectively). Incidentally, this is also the case in Model 3b (which test the effect on marrying one year later) although not statistically precise. In Model 5b, the interaction coefficient becomes negative, suggesting that participating in the military service during a war significantly lowers the probability of marrying three years later with respect to the general population. In this last model, in particular, living in a country at war yields a negative effect, as in the previous regressions, but loses statistical significance.

The regressions on divorce (Table 2) show that serving in the army significantly lowers the probability to divorce, before and after such event (Model 6a, 8a, 9a and 10a) with respect to the general population. Again, the explanation might lie in the fact that participating in the army could promote traditional family views. The effect of living in a country at war is systematically negative too, but never significant.

When we add the interaction variable between living in a country at war and serving in the army, we find that, for those who served in times of peace, the probability of divorcing is significantly lower in the years following the conscription (Models 8b, 9b and 10b). However, and interestingly, serving in times of conflict is associated to a probability of divorcing later on that rises consistently over time and that is always

Table 1. Regressions on marrying at various years ($t-1$ to $t+3$)
Linear probability model, standard error in parentheses. Controlled for previous marital status, age, whether the respondent had a kid and economic conditions.

	$t-1$		t		$t+1$		$t+2$		$t+3$	
	Model 1a	Model 1b	Model 2a	Model 2b	Model 3a	Model 3b	Model 4a	Model 4b	Model 5a	Model 5b
Serving in the army	-.0120*** (.0033)	-.0140*** (.0041)	.0005 (.0031)	-.0016 (.0038)	.0355*** (.0032)	.0352*** (.0039)	.0612*** (.0033)	.0555*** (.0040)	.0676*** (.0034)	.0732*** (.0042)
Living in a country at war	-.0047*** (.0020)	-.0052*** (.0021)	-.0042*** (.0019)	-.0048*** (.0020)	-.0037* (.0020)	-.0038* (.0021)	-.0032 (.0020)	-.0048*** (.0021)	-.0040* (.0021)	-.0024 (.0022)
Serving in the army # living in a country at war		.0057 (.0070)		.0069 (.0067)		.0009 (.0069)		.0169** (.0070)		-.0164*** (.0071)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2. Regressions on divorcing at various years ($t-1$ to $t+3$)
Linear probability model, standard error in parentheses. Controlled for previous marital status, age, whether the respondent had a kid and economic conditions.

	$t-1$		t		$t+1$		$t+2$		$t+3$	
	Model 6a	Model 6b	Model 7a	Model 7b	Model 8a	Model 8b	Model 9a	Model 9b	Model 10a	Model 10b
Serving in the army	-.0021** (.0008)	-.0022** (.0011)	-.0011 (.0008)	-.0008 (.0010)	-.0021** (.0008)	-.0020* (.0010)	-.0017* (.0009)	-.0024** (.0011)	-.0022** (.0009)	-.0034*** (.0011)
Living in a country at war	-.0001 (.0005)	-.0001 (.0005)	-.0005 (.0005)	-.0005 (.0005)	-.0004 (.0005)	-.0004 (.0005)	-.0005 (.0005)	-.0007 (.0006)	-.0004 (.0006)	-.0008 (.0006)
Serving in the army # living in a country at war		.0001 (.0018)		-.0009 (.0018)		-.0001 (.0018)		.0021 (.0019)		.0037* (.0020)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

lower than for those who served in time of peace (from Model 8b onwards). In our last regression (Model 10b, testing the effect on divorcing three years later) the coefficient is positive and statistically significant.

Altogether, these results seem to highlight that, in the post-Soviet era, participating in the military service was associated to higher chances of marrying and lower chances of divorcing compared to the general population. However, these effects are at least mitigated for those who have served specifically during a time of war and might have had, therefore, a direct involvement in the armed conflict: in this case, the marital stability of army members seems challenged. This is coherent with what the existing literature on other cultural and socio-demographic contexts has shown.

Acknowledgements. This work has been published as part of the project UnFamiLiC – Uncertainty and the Family Life Course (CUP B55F21007810001), funded by NextGen EU funds and supported by the Ministry of Education, University and Research.

References

- Anderson, W.L., Little, D.W.: All's fair.: war and other causes of divorce from a beckerian perspective. *Am. J. Econ. Sociol.* **58**(4), 901–922 (1999)
- Angrist, J.D., Johnson, J.H.: Effects of work-related absences on families: evidence from the gulf war. *ILR Rev.* **54**(1), 41–58 (2000)
- Cesur, R., Sabia, J.J., Tekin, E.: The psychological costs of war: military combat and mental health. *J. Health Econ.* **32**(1), 51–65 (2013)
- Cook, J.M., Riggs, D.S., Thompson, R., Coyne, J.C., Sheikh, J.I.: Posttraumatic stress disorder and current relationship functioning among World War II ex-prisoners of war. *J. Fam. Psychol.* **18**(1), 36 (2004)
- Janmardy, M.: Study of divorce in Iran provinces from 1977 to 1998: emphasis on the role of Iran-Iraq war. *Int. J. Sociol. Anthropol.* **3**(4), 132–138 (2011)
- Lester, D.: The effect of war on marriage, divorce and birth rates. *J. Divorce Remarriage* **19**(1–2), 229–232 (1993)
- Negrusa, S., Negrusa, B., Hosek, J.: Gone to war: have deployments increased divorces? *J. Popul. Econ.* **27**(2), 473–496 (2014)
- Pavalko, E.K., Elder, G.H.: World War II and divorce: a life-course perspective. *Am. J. Sociol.* **95**(5), 1213–1234 (1990)
- Resetar, B., Berdica, J.: Divorce in Croatia: the principles of no-fault divorce, parental responsibility, parental education, and children's rights. *Fam. Ct. Rev.* **51**, 568 (2013)
- Zelder, M.: Inefficient dissolutions as a consequence of public goods: the case of no-fault divorce. *J. Leg. Stud.* **22**(2), 503–520 (1993)



Bayesian Inference for Discrete Spatio-Temporal Hidden Markov Models

Daniele Tancini¹✉, Riccardo Rastelli², and Francesco Bartolucci¹

¹ University of Perugia, Perugia, Italy

daniele.tancini@dottorandi.unipg.it, francesco.bartolucci@unipg.it

² University College Dublin, Dublin, Ireland

riccardo.rastelli@ucd.ie

Abstract. Spatio-temporal hidden Markov models are extremely difficult to estimate because their latent joint distributions are available only in trivial cases. These distributions are usually substituted with pseudo-distributions and this could affect the estimation results, in particular with strong dependence between the latent variables. In this work, we show how inference can be carried out in a Bayesian framework using a Markov chain Monte Carlo (MCMC) algorithm, which eliminates the need to calculate the distribution of the latent variables. This approach is based on the exchange algorithm, an MCMC method which is extensively used for doubly intractable likelihood problems. Since the use of this algorithm is rather uncommon for spatio-temporal models, we propose here an application to this context and provide a simulation study where we compare the pseudo-likelihood MCMC and the proposed algorithm.

Keywords: Exchange algorithm · intractable likelihood · Markov chain Monte Carlo · pseudo-likelihood

1 Introduction

Spatio-temporal hidden Markov models are important in many research areas. Unfortunately, the intractability of the distribution of the latent variable is a problem for the estimation of these models. For this reason, in Bayesian inferential methods the likelihood of the unobservable component is substituted with a pseudo-likelihood; see [1,2]. However, this approximation could affect the estimation results, presumably when the model complexity is high.

A possible solution to the problem described above is the exchange algorithm [3], which has been extensively used in exponential random graph models [4] and in general for doubly intractable likelihood problems. For a complete review on Bayesian inference in the presence of intractable normalizing constants see [5].

In this work, we show how the exchange algorithm can be used for spatio-temporal hidden Markov models. In particular, the proposed algorithm is an exchange algorithm, where also the latent component is augmented in the Markov chain Monte Carlo (MCMC) scheme.

The present work is organized as follows. An outline of spatio-temporal hidden Markov models and the exchange MCMC algorithm is provided in Sect. 2. In Sect. 3 we present a comparison between the pseudo-likelihood MCMC and the proposed algorithm.

2 Model

In a spatio-temporal context, let $\mathcal{T} = \{1, \dots, T\}$ be the set of time occasions and $\mathcal{S} = \{1, \dots, N\}$ the site set; we consider a pair of stochastic processes $\{\mathbf{Y}_{it}, U_{it} : i \in \mathcal{S}, t \in \mathcal{T}\}$, where $\mathbf{Y}_{it}$ is observable and U_{it} is latent. We define with $\mathcal{U}$ the discrete state space of the latent component.

We assume for $\{U_{it}\}$ a first order Markov (time) property combined with a Markov random (space) property with respect to a neighborhood system $\mathcal{H}$, defined as $\mathcal{H} = \{\eta_i : i \in \mathcal{S}\}$, with η_i being the neighborhood of site i such that $i \notin \eta_i$ and if $j \in \eta_i$ then $i \in \eta_j$. A typical assumption on these models is that of local independence, meaning that the observable variables are conditionally independent given the latent variables. A Gaussian spatio-temporal hidden Markov model assumes that

$$\mathbf{Y}_{it}|U_{it} = u \sim \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u), \quad i = 1, \dots, N, t = 1, \dots, T.$$

Notice that different distributions can be considered instead of the Gaussian one and it is also possible to include covariates.

For the latent component, we consider the following representation based on a suitable vector of parameters $\boldsymbol{\theta}$:

$$p(\mathbf{u}|\boldsymbol{\theta}) = \frac{q_{\boldsymbol{\theta}}(\mathbf{u})}{\mathcal{Z}_{\boldsymbol{\theta}}},$$

where $q_{\boldsymbol{\theta}}(\mathbf{u}) = \exp(\mathbf{f}(\mathbf{u})'\boldsymbol{\theta})$ and $\mathcal{Z}_{\boldsymbol{\theta}} = \sum_{\mathbf{u}} q_{\boldsymbol{\theta}}(\mathbf{u})$, with $\mathbf{f}(\mathbf{u})$ being a vector which summarizes the states of the neighbors of site i at occasion t and accounts for the state at time $t - 1$. Notice that the normalizing constant $\mathcal{Z}_{\boldsymbol{\theta}}$ is extremely difficult to evaluate, usually intractable, since the sum is extended to all possible latent configurations $\mathbf{u}$.

In this work, we extend the autologistic model [2] by including time dependence as follows:

$$\begin{aligned} \log q_{\boldsymbol{\theta}}(\mathbf{u}) = & \sum_{i=1}^N u_{i1}\beta + \sum_{i=1}^{N-1} \sum_{j>i:j \in \eta_i} u_{i1}u_{j1}\gamma + \sum_{t>1} \left(\sum_{i=1}^N u_{it}\beta^* + \right. \\ & \left. + \sum_{i=1}^{N-1} \sum_{j>i:j \in \eta_i} u_{it}u_{jt}\gamma^* + \sum_{i=1}^N u_{it}u_{it-1}\delta \right), \end{aligned}$$

where $\boldsymbol{\theta} = (\beta, \beta^*, \gamma, \gamma^*, \delta)'$ and $\mathcal{U} = \{-1, 1\}$ is the support of each variable. Assuming independence *a priori* between parameters, we can write the posterior distribution as

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta} | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) p(\boldsymbol{\gamma}) p(\boldsymbol{\delta}) p(\boldsymbol{\beta}), \quad (1)$$

where

$$p(\mathbf{y}|\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{u \in \mathcal{U}} \prod_{i \in \mathcal{S}} \prod_{t \in \mathcal{T}} e^{-\frac{1}{2}(\mathbf{y}_{it} - \boldsymbol{\mu}_u)' \boldsymbol{\Sigma}_u^{-1} (\mathbf{y}_{it} - \boldsymbol{\mu}_u) \mathbb{1}_{U_{it}}(u)},$$

with $\mathbb{1}_{U_{it}}(u)$ as indicator function, which takes value 1 if $U_{it} = u$ and 0 otherwise.

It is easy to verify that a classical MCMC algorithm cannot be used in this case, since the normalizing constant $\mathcal{Z}_{\boldsymbol{\theta}}$ depends on $\boldsymbol{\theta}$ and does not simplify. A possible solution is the exchange algorithm [3]. This algorithm samples from an augmented distribution, where an auxiliary variable $\boldsymbol{\omega}$, which lives in the same space of the latent $\mathbf{u}$, is included. Following this approach, it is possible to simplify the normalizing constant; see [3]. In order to estimate this model we consider an MCMC algorithm combining exchange and Gibbs steps.

We propose the following prior distributions for the parameters involved in the conditional distribution of the responses:

$$\boldsymbol{\mu}_u \sim \mathcal{N}(\mathbf{m}, \mathbf{V}), \quad \boldsymbol{\Sigma}_u \sim \mathcal{IW}(\nu, \mathbf{S}), \quad u = -1, 1,$$

where $\mathcal{IW}(\cdot, \cdot)$ denoted an Inverse-Wishart distribution, while for the latent distribution we have

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \cdot \mathbf{I}), \quad \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\gamma}}^2 \cdot \mathbf{I}), \quad \delta \sim \mathcal{N}(0, \sigma_{\delta}^2).$$

From Eq. (1), we can obtain the following full conditionals. For the mean vectors we have

$$\boldsymbol{\mu}_u | \cdots \sim \mathcal{N}(\tilde{\mathbf{V}}_u \tilde{\mathbf{m}}_u, \tilde{\mathbf{V}}_u),$$

where

$$\tilde{\mathbf{V}}_u^{-1} = n_u \boldsymbol{\Sigma}_u^{-1} + \mathbf{V}^{-1} \quad \text{and} \quad \tilde{\mathbf{m}}_u = \boldsymbol{\Sigma}_u^{-1} n_u \bar{\mathbf{y}}_u + \mathbf{V}^{-1} \mathbf{m},$$

with

$$n_u = \sum_{i=1}^N \sum_{t=1}^T \mathbb{1}_{U_{it}}(u) \quad \text{and} \quad \bar{\mathbf{y}}_u = (1/n_u) \sum_{i=1}^N \sum_{t=1}^T \mathbf{y}_{it} \mathbb{1}_{U_{it}}(u).$$

For the variance-covariance matrices we have

$$\boldsymbol{\Sigma}_u | \cdots \sim \mathcal{IW}(\nu + n_u, \mathbf{S} + \tilde{\mathbf{S}}_u),$$

where

$$\tilde{\mathbf{S}}_u = \sum_{i=1}^N \sum_{t=1}^T (\mathbf{y}_{it} - \boldsymbol{\mu}_u)(\mathbf{y}_{it} - \boldsymbol{\mu}_u)' \mathbb{1}_{U_{it}}(u).$$

These full conditionals can be obtained following the same approach as in [6]. For $\boldsymbol{\beta}$ we propose the following step. Generate a new $\tilde{\boldsymbol{\beta}}$ using a symmetric proposal $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I})$, and then accept this value with probability

$$\alpha(\beta, \tilde{\beta}) = 1 \wedge \frac{p(\tilde{\beta})q(\mathbf{u}; \tilde{\beta}, \delta^{(r)}, \gamma^{(r)})q(\boldsymbol{\omega}; \beta^{(r)}, \delta^{(r)}, \gamma^{(r)})}{p(\beta^{(r)})q(\mathbf{u}; \beta^{(r)}, \delta^{(r)}, \gamma^{(r)})q(\boldsymbol{\omega}; \tilde{\beta}, \delta^{(r)}, \gamma^{(r)})}, \quad (2)$$

with $q(\mathbf{u}; \tilde{\beta}, \delta^{(r)}, \gamma^{(r)}) = q_{\tilde{\theta}}(\mathbf{u})$ and $\boldsymbol{\omega} \sim q_{\tilde{\theta}}(\cdot)/Z_{\tilde{\theta}}$. The same step is followed for γ and δ . In this case, we need to draw $\boldsymbol{\omega}$ using a Gibbs sampler, taking the last iteration for the evaluation of the ratio in Eq. (2). Theoretical justifications for using the final sample can be found in [7]. Finally, we have a Gibbs step for $\{U_{it}\}$; in particular, when $1 < t < T$ we have

$$p(U_{it} = u | \dots) \propto \exp \left(\beta^* u + \sum_{j>i: j \in \eta_i} uu_{jt} \gamma^* + uu_{it+1} \delta + uu_{it-1} \delta \right) \\ \times p(\mathbf{y}_{it} | U_{it} = u; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u),$$

while different forms are provided for $t = 1$ and $t = T$.

3 Simulation Study

In this section we present a simulation study where we compare the algorithm proposed and the pseudo-likelihood MCMC approach. The latter, is simply obtained using for $\{U_{it}\}$ the pseudo-likelihood, meaning that the exchange steps are substituted by Metropolis-Hastings steps.

We consider a simulation design where $N = 50$ and $T = 5, 20$. For both algorithms we consider the same starting points and number of iterations. In particular we fix as 100,000 the number of MCMC iterations and we consider 50% of burn-in. The hyperparameters for the priors are selected in order to provide minimal information on the parameters. We evaluate the results on the basis of conditional expectations, Monte Carlo standard errors (MCSE) calculated by using the batch means estimator [8] and the $\hat{R}$ proposed by [9] and implemented in the **rstan** package [10]. These results are reported in Table 1.

For $T = 5, 20$, the means and variance-covariance matrices are well estimated by both algorithms, obtaining small values of MCSE and for all the parameters $\hat{R} < 1.05$, as suggested in [10]. Regarding the parameters of $\{U_{it}\}$, we can note that the MCSE for the exchange algorithm are quite smaller than for the pseudo-likelihood algorithm and, in general, the estimates obtained by the first are more precise than those obtained by the second, especially when $T = 20$. These results could be explained given the approximation provided by the pseudo-likelihood, leading to imprecise estimates when the model complexity increases. Both algorithms present for all the parameters an $\hat{R} < 1.05$. These results are summarized in Figures 1 and 2.

We conclude with a comparison on the computational time. For the pseudo-MCMC are required 9.582 milliseconds for each iteration, while 25.025 millisec-

Table 1. Comparison between exchange and pseudo-likelihood MCMC algorithms.

State	Parameters	Estimate (Monte Carlo s.e.)		True	$\hat{R}$	
		Exchange	Pseudo-lik.		Exchange	Pseudo-lik.
T = 5						
-1	μ_1	-0.939 (0.001)	-0.940 (0.001)	-1.0	1.000	1.000
	μ_2	-1.963 (0.001)	-1.963 (0.001)	-2.0	1.000	1.000
	$\sigma_{1,1}$	1.096 (0.001)	1.097 (0.001)	1.0	1.000	1.000
	$\sigma_{1,2}$	-0.605 (0.001)	-0.606 (0.001)	-0.6	1.000	1.000
	$\sigma_{2,2}$	0.992 (0.001)	0.993 (0.001)	1.0	1.000	1.000
1	μ_1	1.969 (0.001)	1.970 (0.001)	2.0	1.000	1.000
	μ_2	4.065 (0.001)	4.065 (0.001)	4.0	1.000	1.000
	$\sigma_{1,1}$	1.131 (0.001)	1.132 (0.001)	1.0	1.000	1.000
	$\sigma_{1,2}$	0.905 (0.001)	0.907 (0.001)	0.8	1.000	1.000
	$\sigma_{2,2}$	1.037 (0.001)	1.040 (0.001)	1.0	1.000	1.000
	β	-0.555 (0.185)	0.430 (0.389)	-1.0	1.001	1.001
	β^*	-0.253 (0.179)	-0.253 (0.280)	2.0	1.002	1.017
	γ	-2.288 (0.171)	-5.462 (0.141)	-2.0	1.001	1.002
	γ^*	-2.416 (0.146)	-2.221 (0.173)	-4.0	1.000	1.005
	δ	-5.072 (0.134)	-6.249 (0.154)	-3.0	1.001	1.000
T = 20						
-1	μ_1	-1.011 (0.001)	-1.011 (0.001)	-1.0	1.000	1.000
	μ_2	-1.991 (0.001)	-1.992 (0.001)	-2.0	1.000	1.000
	$\sigma_{1,1}$	0.998 (0.001)	0.998 (0.001)	1.0	1.000	1.000
	$\sigma_{1,2}$	-0.589 (0.001)	-0.589 (0.001)	-0.6	1.000	1.000
	$\sigma_{2,2}$	1.045 (0.001)	1.046 (0.001)	1.0	1.000	1.000
1	μ_1	1.991 (0.001)	1.991 (0.001)	2.0	1.000	1.000
	μ_2	4.009 (0.001)	4.009 (0.001)	4.0	1.000	1.000
	$\sigma_{1,1}$	1.070 (0.001)	1.072 (0.001)	1.0	1.000	1.000
	$\sigma_{1,2}$	0.871 (0.001)	0.873 (0.001)	0.8	1.000	1.000
	$\sigma_{2,2}$	1.112 (0.001)	1.113 (0.001)	1.0	1.000	1.000
	β	-2.550 (0.172)	0.089 (0.227)	-1.0	1.026	1.000
	β^*	1.697 (0.036)	0.295 (0.326)	2.0	1.002	1.008
	γ	-3.096 (0.166)	-5.707 (0.162)	-2.0	1.002	1.000
	γ^*	-5.154 (0.090)	-9.951 (0.244)	-4.0	1.001	1.013
	δ	-3.958 (0.095)	-4.010 (0.134)	-3.0	1.021	1.008

onds are necessary for the exchange algorithm. This difference is given by the auxiliary variables (for each parameter) required in the exchange MCMC, which are sampled using a Gibbs sampler.

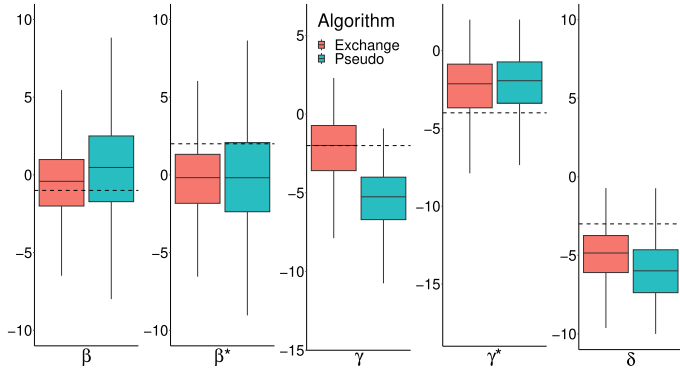


Fig. 1. Simulation study for the case $T = 5$. The boxplots represent the posterior samples for $\theta = (\beta, \beta^*, \gamma, \gamma^*, \delta)'$ for the exchange and the pseudo-likelihood MCMC algorithms. The horizontal dashed lines show the true parameter value.

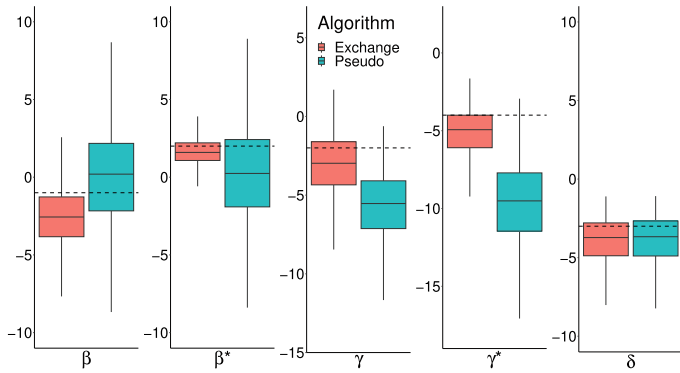


Fig. 2. Simulation study for the case $T = 20$. The boxplots represent the posterior samples for $\theta = (\beta, \beta^*, \gamma, \gamma^*, \delta)'$ for the exchange and the pseudo-likelihood MCMC algorithms. The horizontal dashed lines show the true parameter value.

Acknowledgments. The authors acknowledge the financial support from the grant “Latent variable models: methodological and applicative developments for lifestyle and society” of the University of Perugia (University Research Fund - FRA 2022 - Working Package 6.2).

References

1. Bartolucci, F., Farcomeni, A.: A hidden Markov space-time model for mapping the dynamics of global access to food. *J. R. Stat. Soc. Ser. A* **185**, 246–266 (2022)
2. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. B* **36**, 192–225 (1974)
3. Murray, I., Ghahramani, Z., MacKay, D.: MCMC for doubly-intractable distributions. arXiv preprint [arXiv:1206.6848](https://arxiv.org/abs/1206.6848) (2012)

4. Caimo, A., Friel, N.: Bayesian inference for exponential random graph models. *Soc. Netw.* **33**, 41–55 (2011)
5. Park, J., Haran, M.: Bayesian inference in the presence of intractable normalizing functions. *J. Am. Stat. Assoc.* **113**, 1372–1390 (2018)
6. Tancini, D., Bartolucci, F., Pandolfi, S.: A comparison between marginal likelihood and data augmented MCMC algorithms for Gaussian hidden Markov models. *J. Stat. Comput. Simul.* **94**, 1571–1594 (2024)
7. Everitt, R.G.: Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comput. Graph. Stat.* **21**, 940–960 (2012)
8. Flegal, J.M., Galin, L.J.: Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Stat.* **38**, 1034–1070 (2010)
9. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., Burkner, P.C.: Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16**, 667–718 (2021)
10. Stan Development Team. RStan: the R interface to Stan. R package version 2.32.5 (2024). <https://mc-stan.org/>



Modelling Three-Dimensional Point Patterns

Marco Tarantino^(✉), Nicoletta D'Angelo, and Giada Adelfio

Department of Business, Economics and Statistics, University of Palermo, Viale delle Scienze 8, Palermo, PA, Italy
`marco.tarantino01@unipa.it`

Abstract. In this study, we explore the cubature scheme procedure for modelling three-dimensional point patterns through Poisson point process models, a computational realm that remains under-explored. Through simulations, we give guidelines for choosing the number of cubes to partition the observed three-dimensional region, the number of dummy points to generate, and whether to simulate them regularly or casually in space. We apply this methodology to the real observed point pattern of young stars of the Gaia Archive.

Keywords: three dimensional point processes · initial parameter tuning · model implementation

1 Introduction

Patterns of points in three-dimensional space, as well as those occurring in both space and time, have growing attention in data description methods, since the occurrence of natural processes in three-dimensional space and their evolution over time. Three-dimensional geometry is particularly important in geology, anatomy and astronomy [1]. Examples of spatio-temporal point patterns are available in epidemiology (e.g., for disease surveillance), seismology, anatomy (e.g., video microscopy), and increasingly in astronomy, where new telescopes make it possible to observe transient events across a wide survey volume [3].

In this paper, we explore the cubature scheme procedure for modelling three-dimensional point patterns through Poisson point process models, improving the parameters estimation of the linear predictor, when specific covariates are available. Specifically, the study is based on the available cubature scheme approaches proposed in the package *stopp* [4], which extends the most known purely spatial quadrature scheme of *spatstat* [2] to the spatio-temporal context. The main goal of this paper is to assess the estimation capability of the Poisson point process model in the context of three-dimensional point processes. Specifically, we investigate how the model's performance is affected by variations in some initial parameters, looking for the optimal values for these parameters by simulations. In particular, we considered data simulated from homogeneous Poisson processes and non-homogeneous Poisson processes, including the dependence

on spatial coordinates. The simulated scenarios depend on the specification of three prior choices: 1) q : the multiplier of the data points to choose the number of dummy points to generate; 2) n_c : number of cubes into which the observed three-dimensional space is partitioned to define the cubature scheme; 3) *grid*: the type of grid, specifying whether the dummy points are placed regularly in the space or randomly. Simulations are performed on different combinations of the three choices, henceforth called initial parameters. Through the conducted simulations, we found specific values for the initial parameters that resulted in a significant reduction in Mean Squared Error (MSE). On the basis of these results, the model is applied to a real data set, concerning young stars, for which the galactic coordinates are available in the Gaia Archive [5], transformed into spatial coordinates, coherently to a three-dimensional point pattern. All the analyses are carried out through the R statistical software and are available from the authors.

The structure of the paper is as follows. In Sect. 2 we illustrate the proposed methodology to fit three-dimensional Poisson processes. Then, Sect. 3 contains the results of the simulations, showing the best values of *grid*, q and n_c for each scenario. Furthermore, in Sect. 4 we report one application of the function on the stellar dataset, considering the stars as a point process in three dimensions. The paper ends with conclusions and possible future works in Sect. 5.

2 Methodology

As previously stated, our emphasis is on Poisson processes in three dimensions. The statistical approach used for two-dimensional point processes [2] can easily be extended to three-dimensional point processes, without significant hurdles. Let X be a three-dimensional point process and $\mathbf{x}$ the observed point pattern in an observed window $W \subset \mathbb{R}^3$; its parametric intensity, or rate function, can be defined as $\lambda(u; \theta) = \exp(\theta^T \mathbf{Z}(u))$ with $u \in W$ denoting the spatial coordinates (x, y, z) , and some spatial covariates $\mathbf{Z}(u)$ with unknown parameters $\theta \in \Theta$. In particular, in this work we consider only the coordinates x , y and z as spatial covariates, available over the entire observed three-dimensional space.

The log-likelihood is

$$\log L(\theta) = \sum_i \lambda(u_i; \theta) - \int_W \lambda(u; \theta) du \quad (1)$$

where the sum is over all points u_i in the point pattern $\mathbf{x}$. For estimation purposes, we use a finite quadrature approximation of the log-likelihood, that is, starting from the observed points $u_1, \dots, u_n$, generate m additional *dummy points* $u_{n+1}, \dots, u_{n+m}$ to form a set of $n + m$ quadrature points (where $m > n$). Then, integral in (1) is approximated by the Riemann sum

$$\int_W \lambda(u; \theta) du \approx \sum_{k=1}^{n+m} a_k \lambda(u_k; \theta)$$

where $a_1, \dots, a_m$ are quadrature weights and $\sum_{k=1}^{n+m} a_k = l(W)$, with $l(\cdot)$ the Lebesgue measure. In order to distinguish between dummy points and observed points, we introduce the indicator e_k , which assumes values equal to 1 if the related point is a data point, zero otherwise. Then, writing $y_k = e_k/a_k$, the log-likelihood (1) of the template model can be approximated by

$$\log L(\theta) \approx \sum_j a_k (y_k \log \lambda(u_k; \theta) - \lambda(u_k; \theta)) + \sum_k a_k.$$

Apart from the constant $\sum_k a_k$, this expression is formally equivalent to the weighted log-likelihood of a Poisson regression model with responses y_k and mean $\lambda(u_k; \theta) = \exp(\theta)$. In such a way, the model can be easily maximised using standard Generalized Linear Model (GLM) approaches. The three-dimensional quadrature scheme is obtained by defying a three-dimensional partition of W into n_c cubes C_k of equal volume ν , assigning the weight $a_k = \nu/n_k$ to each quadrature point (dummy or data), where n_k is the number of points that lie in the same cube as the point u_k . The number of dummy points should be enough for ensuring accurate estimates. Starting with $m \approx 4n$ dummy points, up to a value such that $\sum_k a_k = l(W)$, we have $m = qn$.

3 Simulations Results

We simulate 100 spatial point patterns from three different purely spatial point processes in the unit cube with $n \in \{100, 250, 500\}$ points on average following these three scenarios:

1. homogeneous Poisson process with constant intensity $\lambda = \exp(\theta_0)$,
2. inhomogeneous Poisson process with intensity $\lambda(x, y, z) = \exp(\theta_1 x)$,
3. inhomogeneous Poisson process with intensity $\lambda(x, y, z) = \exp(\theta_0 + \theta_1 x + \theta_2 y + \theta_3 z)$

Different results are compared in terms of MSE, fitting as many models as the number of combinations of the initial parameters. In particular, we consider both logical values of *grid*, $1 \leq q \leq 30$ and $1 \leq n_c \leq 30$. For scenario 3, we explore n_c values up to 45. Given the different number of parameters p to estimate, we consider a measure of overall error for the estimated set of parameters as $MSE = \frac{1}{p} \sum_{j=1}^p MSE(\hat{\theta}_j)$.

Once computed the MSE with respect to the parameters estimated by the models, an average across the 100 simulated point patterns is computed. Once the grid of MSE values is obtained, a search for the values of *grid*, q , and n_c at the minimum error value is then conducted.

In the application on real data presented in Sect. 4, we used all three coordinates as covariates, that is the most complex scenario among those considered. Therefore, in Fig. 1 the graphs showing the MSEs obtained in scenario 3 with three coordinates are reported. The patterns' size is set to 100, 250, and 500, exclusively for the model with a random *grid*, since, in almost all the explored

scenarios, it resulted in lower error values compared to a regular arrangement of dummy points in three-dimensional space. To identify the absolute minimum, a blue dot is included in the graphs, that stands in the combination of q and n_c that yields the lowest observed error (MSE_{min}). Given that a higher value of q leads to the generation of more dummy points and requires more computational time, it is essential to explore values of the MSE not too far from the observed minimum.

Looking for lower values for the initial parameter q , two additional points are displayed on the grids. A green dot corresponds to the ‘minimum’ combination of q and n_c , yielding a MSE value between MSE_{min} and $MSE_{min} + 10\%MSE_{min}$. A red point is assigned to the ‘minimum’ pair of q and n_c , resulting in an MSE value between MSE_{min} and $MSE_{min} + 20\%MSE_{min}$.

Observing the graphs, the point size is proportional to the value of the error; it becomes evident that even considering a 20% higher MSE value than the minimum MSE, it remains very low, comparable to the absolute minimum value within the grid.

This result suggests the use of significantly fewer dummy points than necessary to achieve the minimum MSE, without compromising the accuracy of model estimation and leading to substantial computational time savings. In addition, the black line in Fig. 1 shows the ‘default’ values n_c and q corresponding to the minimum MSE for the simplest scenarios 1 and 2. It is evident that for scenario 3, the black lines deviate significantly from the white areas of the grid, indicating the lowest MSE value areas.

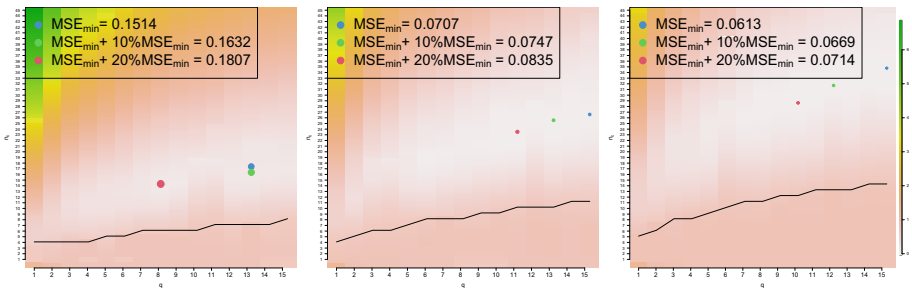


Fig. 1. MSE values obtained for the inhomogeneous Poisson processes with dependence on all three coordinates and sizes 100, 250 and 500 (scenario 3). On the x-axis are the values of q , on the y-axis are the values of n_c . The error scale is the same for all the three grids. Point size is proportional to MSE.

4 Application on Young Stars Data

Once obtained the best values of the initial parameters $grid$, n_c and q by the simulated scenarios, an application on real three-dimensional spatial data is provided

in this section. Specifically, the data refer to a set of 1169 pre main-sequence stars, also called young stars, analysed as a three dimensional point pattern. For each star, several variables are obtained from the Gaia space mission archive [5]; more specifically, for the application of interest, the galactic coordinates of these stars are extracted. They refer to the celestial coordinate system centered on the Sun and aligned with the center of the Milky Way. Similar to geographic coordinates, galactic coordinates are based on longitude (l) and latitude (b). To obtain the Cartesian coordinates (x , y and z) of these stars, the following mathematical transformations are performed: $x = d \cdot \cos(b)\cos(l)$, $y = d \cdot \cos(b)\sin(l)$ and $z = d \cdot \sin(b)$, where d represents the distance of the star from the Earth (Fig. 2).

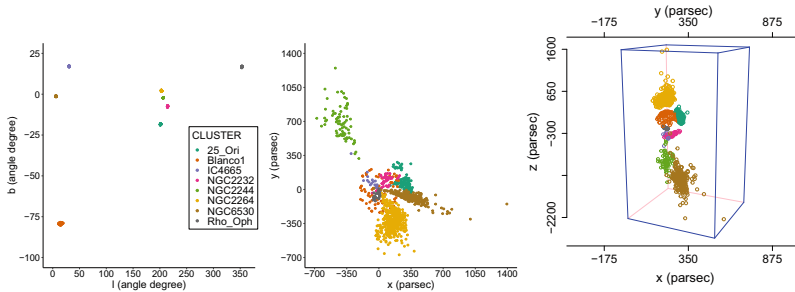


Fig. 2. Spatial representation of the 1169 stars. On the left we consider the space defined by galactic coordinates l and b , in the middle the two-dimensional space in x and y Cartesian coordinates, and on the right the three-dimensional space in x , y and z Cartesian coordinates.

Analysing the dataset, starting from the default values ($q = 4$, $n_c = 13$, and a random *grid*) the coefficients of all the three coordinates turn out to be statistically significant. Therefore, the results obtained from the simulations on the three-coordinate dependent non-homogeneous Poisson process are considered, to suggest possible values of the initial parameters $q, n_c, grid$. Given that the simulations were conducted up to a size of 500, and the reference dataset comprises 1169 stars, we adjust the values of q and n_c accordingly. This procedure is carried out for all three points highlighted in the grids, aiming to fit the model across three different error ranges.

In Table 1, we report the coefficient estimates of the three models' parameters and of the model corresponding to the default q and n_c . As can be seen from Table 1, the intercept is the most variable coefficient estimate across the models' results. Increasing q and n_c , its estimate increases as well. The coefficients associated with the three coordinates exhibit less variability and are consistently close to zero, despite being highly significant. This is likely due to the spatial coordinates' units. We could conclude that any combination of q and n_c obtained from the minimization process consistently enhances the estimates compared to the default. Moreover, among the three minima of interest, there

are only slight changes in the MSE. Consequently, it is possible to stop at a lower computational effort ($q = 10$) compared to what is required to achieve the absolute minimum ($q = 16$), albeit the number of dummy points generated is higher than the number generated by default ($q = 4$).

Table 1. Estimates of coefficients obtained from the models for some specific combinations of q and n_c .

Models	Intercept	x	y	z	q	n_c	MSE
default	-15.925	-0.000565	-0.00133	0.000119	4	13	0.8012
$MSE_{min} + 20\%MSE_{min}$	-14.794	-0.000563	-0.00128	0.000118	10	34	0.0714
$MSE_{min} + 10\%MSE_{min}$	-14.667	-0.000573	-0.00132	0.000115	12	38	0.0669
MSE min	-14.586	-0.000555	-0.00133	0.000116	16	43	0.0613

5 Conclusion

The objective of this study was to investigate the cubature scheme procedure for modelling three-dimensional point processes, identifying optimal values for the three initial parameters associated with low estimation errors, in terms of MSE. A simulation plan was carried out to analyse the grids of MSE values corresponding to different scenarios and initial parameter values. The results indicate that the default values of q and n_c established by existing literature may not be optimal for parameter estimates, depending on the considered scenario. Additionally, a new R function was implemented, based on functions available in the *stopp* and *spatstat* packages, to model three-dimensional point processes. This function allows users to specify the initial values of the three parameters (q , n_c , and *grid*). The implemented function can be straightforwardly extended to model marked point processes, also incorporating external continuous spatial covariates, such as environmental variables, into the model. Future work will then involve simulations to explore optimal initial parameter values for these extended scenarios.

Fundings. This work was supported by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C93C22005270001).

References

1. Babu, G.J., Feigelson, E.D.: Spatial point processes in astronomy. *J. Stat. Plann. Inference* **50**(3), 311–326 (1996)

2. Baddeley, A., Turner, R.: spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.* **12**(6), 1–42 (2005)

3. Baddeley, A., Turner, R.: Spatial point patterns: methodology and applications with R (2015)
4. D'Angelo, N., Adelfio, G.: stopp: methods for spatio-temporal point pattern analysis, simulation, model fitting, diagnostics, and local analyses. R package version 0.1.0 (2023)
5. Gaia Archive. <https://gea.esac.esa.int/archive/>



Presmoothing Methodology for Instrumental Variable Estimation of the Proportional Odds Model Under Random Right Censoring

Lorenzo Tedesco^(✉)

University of Bergamo, Bergamo, Italy
lorenzo.tedesco@unibg.it

Abstract. This study explores the instrumental variable estimation within the framework of the semi-parametric proportional odds model. The analysis accommodates both discrete instrumental and endogenous variables, alongside potentially continuous exogenous covariates. By employing a rank invariance hypothesis, it is possible to reformulate the proportional odds model into a semiparametric version of the instrumental variable quantile regression model of [3]. We based the estimation on the presmoothing methodology proposed in [10] which is composed of three steps. First, we estimate the model nonparametrically - and show that this nonparametric estimator has a closed-form solution in the leading case of interest of randomized experiments with one-sided noncompliance. Second, we use the nonparametric estimator to generate “proxy” observations for which exogeneity holds. Third, we apply the usual maximum likelihood estimator to the “proxy” data. The estimation procedure allows for random right-censoring. The approach is illustrated via simulation studies.

Keywords: Proportional odds model · Instrumental variable · Presmoothing · Nonparametric estimation

1 Introduction

This paper delves into the semi-parametric Proportional Odds (PO) model, focusing on time-to-event data under conditions of random right-censoring. Given the prevalent challenge of unmeasured confounding in observational studies-where such confounders affect both treatment and outcome, leading to endogeneity-we explore solutions through an instrumental variable (IV) approach.

Addressing endogeneity within the PO framework, we align with the methodology outlined by [10]. That is, the semiparametric PO model is embedded within the instrumental variable quantile regression (IVQR) model as conceptualized by [3], where the relative rank invariance hypothesis is assumed. By constraining the treatment effects to only vary by quantile, this approach allows estimating quantile treatment effects over the full population. This is in contrast to the

local average treatment effects framework of [1], which does not restrict the heterogeneity of treatment effects but requires a monotonicity assumption to hold and only allows to estimate treatment effects on subpopulations. See [11] for a detailed comparison of the two approaches.

As in [10], we specifically address scenarios with discrete endogenous variables and instruments assuming the same number of modalities, alongside (potentially continuous) exogenous covariates. This scenario is particularly relevant, covering a spectrum from randomized trials with noncompliance to more complex experimental designs. Given the focus on time-to-event outcomes with conditional independence in censoring, the method consists of three steps:

1. Estimating the causal quantile regression function nonparametrically.
2. Generating “proxy” observations from this function, where treatment effects are considered exogenous.
3. Applying [9]’s estimation approach to these “proxy” observations.

Step 1 is simplified by the fact that the treatment and instrument are discrete, making the inverse problem well-posed. For Step 1, we use [2]’s estimator of the conditional survival function to construct an empirical version of the system of equations generated by the model and then solve the estimated system to obtain the causal regression function of interest. This allows us to deal with random right-censoring, a common feature of duration data. The fact that we can use maximum likelihood estimator in Step 3 is very convenient since the latter has a convex objective function and is implemented in traditional statistical software. In Step 3 we propose to use the method [9] as it was ready implemented in standard software packages, but any other method as [7, 8] or [5], could also be used.

While several works concern IV methods for proportional hazards model, the literature on the PO model is not vast. The few examples are [6, 12] but these methods are different in the purpose as they study the causal treatment effect only on the compilers. Thus, to the best of the author’s knowledge, this paper represents the first attempt of the estimation of the causal treatment effect on the whole population level within the PO framework.

2 Model

We are interested in the effect of an endogenous treatment Z on a non-negative outcome T . We consider a treatment $Z = (Z_1, \dots, Z_{d_Z})^\top$ expressed as a vector of d_Z binary variables and taking L different values $\{z_1, \dots, z_L\} = \mathcal{Z}$. There are some covariables X with compact support $\mathcal{X} \subset \mathbb{R}^{d_X}$. Let $T(z)$ be the potential outcome duration under treatment equal to $z \in \mathcal{Z}$.

Let us assume that the distribution of $T(z)$ given $X = x$ is continuous for all $x \in \mathcal{X}$. We assume that following structural model holds for $T(z)$ given $X = x$:

$$\logit P(T(z) \leq t|x) = \log A_0(t) + (z^\top, x^\top)\beta_0, \quad (1)$$

where $A_0 : \mathbb{R}_+ \mapsto \mathbb{R}_+$ corresponds to the cumulative baseline hazard and $\beta_0 = (\beta_{0,z}, \beta_{0,x})^\top \in \mathbb{R}^{d_Z+d_X}$ is a vector of regression coefficients. We impose the

consistency condition $T = T(Z)$. Note that this implies that $P(T(z) \geq t|x) = \{1 + \Lambda_0(t) \exp((z^\top, x^\top)\beta_0)\}^{-1}$. We assume to observe a categorical instrumental variable W that has the same number of modalities as the treatment. Therefore, W has support $\mathcal{W} = \{w_1, \dots, w_L\} \subset \mathbb{R}^{d_w}$. The duration T is randomly right censored by a random variable C with support in $\mathbb{R}_+$ so that we do not observe T but $Y = \min(T, C)$. The observables are (Y, δ, Z, X, W) , where $\delta = I(T \leq C)$.

Let us also introduce the random variable

$$Q(z) := \{1 + \Lambda_0(T(z)) \exp((z^\top, X^\top)\beta_0)\}^{-1},$$

which is the structural survival function of $T(z)$ given X evaluated at $T(z)$. The random variables $\{Q(z)\}_{z \in \mathcal{Z}}$ can be thought of as the unobserved heterogeneity of the model. Thus, we have $T(z) = H(z, \cdot | X)^{-1}(Q(z))$, with $H(z, t | X) = \{1 + \Lambda_0(t) \exp((z^\top, X^\top)\beta_0)\}^{-1}$. We assume the rank invariance assumption as in [3], that means there exists a random variable Q such that $Q(z) = Q$ for all $z \in \mathcal{Z}$. This implies that for two subjects i and j such that $X_i = X_j$, $T_i(z) > T_j(z)$ we have $Q_i(z) = Q_i > Q_j = Q_j(z)$, which leads to $T_i(z') > T_j(z')$, for all $z, z' \in \mathcal{Z}$. Hence, the rank in the outcome of any two subjects is the same across all potential outcomes. This assumption restricts the heterogeneity of the treatment effects on the duration: the treatment can change the quantiles of the distribution of the potential outcomes but it cannot change the rank that a subject has in this distribution.

We obtain the following results. Thus, following the same procedure of [10] outlined in the Sect. 1, we can define an estimator $\hat{\beta}$ of β_0 .

Theorem 1. (i) Under the assumptions of Theorem 3.2 of [10] the model is identifiable. (ii) Under the assumptions of Theorem 5.1 of [10], it holds $\hat{\beta} - \beta_0 = o_p(1)$. (iii) Under the assumptions of Theorem 5.3 of [10] $\sqrt{n}\{\hat{\beta} - \beta_0\}$ converges weakly to a mean zero Normal random variable.

3 Simulation Results

In this section, we evaluate the finite sample performance of our proposed method using Monte Carlo simulation. Our main objective is to compare the performance of estimator of [9] an estimator that does not take endogeneity into account, with our new estimator. Note that the proposed estimator is once again the estimator of [9] applied after the presmoothing methodology in order to handle the endogeneity issue.

We explore various simulation designs, detailed in Table 1, to thoroughly evaluate our method. For each design, we set $\Lambda_0(s) = s$, and the coefficients $\beta_0 = (\beta_{0,z}, \beta_{0,x})^\top$ as specified in Table 1. It is important to note the endogeneity of variable Z , influenced by the unobserved heterogeneity term U , and the fact that W is a valid instrumental variable.

We model the censoring variable C as an exponential distribution with varying rates to simulate different censoring levels, namely 15% and 30%. Samples of sizes $n = 500$ and 1000 observations are generated to match the distribution of (Y, δ, Z, X, W) , with $Y = \min(T, C)$ and $\delta = I(T \leq C)$, where T represents the event time. The results are based on $N = 1000$ replications.

In this study, we assess the performance of our estimator by presenting the average bias, standard deviation, and Mean Squared Error (MSE) for each component of the parameter vector $\beta = (\beta_z, \beta_x)^\top$. Moreover, we examine the effectiveness of 95% bootstrap confidence intervals in capturing the true parameter values. These intervals are constructed through a naive bootstrap resampling technique with replacement to estimate the standard deviation of our estimator. This estimated standard deviation is then used within a normal approximation to determine the confidence interval bounds. To enhance the efficiency of our simulations, we utilized the acceleration method proposed by [4]. Furthermore, to perform a component-wise evaluation of our estimator, we calculate the Root Mean Squared Error (RMSE) as an indicator of its accuracy. The RMSE is derived using the formula $RMSE = \sqrt{N^{-1} \sum_{j=1}^N \|\hat{\beta}^{(j)} - \beta_0\|^2}$, where $\hat{\beta}^{(j)}$ represents the estimate from the j^{th} simulation. Additional implementation specifics are discussed in detail in [10].

Our findings indicate that the proposed estimator exhibits low bias and, as sample size increases, both bias and coverage rates converge towards their theoretical benchmarks of 0 and 0.95, respectively. The estimator's performance is better with lower proportions of censored observations. Particularly, the discrete Bernoulli design demonstrates superior performance due to its simpler discrete exogenous structure compared to continuous models. Contrary to our method, the estimator of [9] shows bias, underscoring the effectiveness of our approach in handling endogeneity (Table 2).

Table 1. Distribution and coefficient specifications for each design. Note that $B(p)$ stands for the Bernoulli distribution of parameter p , $\mathcal{U}(a, b)$ represents the uniform distribution on the interval $[a, b]$, $\text{Beta}(\alpha, \beta)$ represents the Beta distribution with shape parameters α and β , and ϵ is a standard normal variable. The random variable U is the unobserved heterogeneity of the model, which follows a standard uniform distribution $U \sim \mathcal{U}(0, 1)$.

Design	W	X	Z	$(\beta_{0,z}, \beta_{0,x})$
continuous Beta	$B(0.5)$	$\text{Beta}(2, 5)$	$I(W = 1)I(0.5U - W + 0.75 + 0.5\epsilon \geq 0)$	$(0.7, 0.2)$
continuous uniform	$B(0.5)$	$\mathcal{U}(-0.5, 0.5)$	$I(W = 1)I(0.5U - W + 0.75 + 0.5\epsilon \geq 0)$	$(0.7, 0.2)$
discrete Bernoulli	$B(0.5)$	$B(0.5)$	$I(W = 1)I(0.5U - W + 0.65 + 0.3X + 0.5\epsilon \geq 0)$	$(0.7, 0.7)$

Table 2. Simulation results for each design. We include for the proposed estimator and the estimator in [9], for each level of censoring (Cens) and for different sample sizes (n), the average of the bias (Bias), the standard deviation (Sd) and the Mean Squared Error (MSE) for each component (Comp) of $\beta = (\beta_z, \beta_x)^\top$. For each case, we also include the Root Mean Squared Error value (RMSE). For the proposed estimator, we also provide the coverage of the 95% confidence intervals (CP95) for each component of β .

				proposed estimator					Estimator in [9]			
Design	Cens	n	Comp	Bias	Sd	MSE	RMSE	CP95	Bias	Sd	MSE	RMSE
continuous Beta	15%	500	β_z	−0.036	0.510	0.261	0.839	0.963	−0.512	0.257	0.328	0.730
			β_x	−0.041	1.043	1.090		0.968	−0.003	0.702	0.493	
		1000	β_z	−0.008	0.359	0.129	0.649	0.936	−0.514	0.184	0.298	0.646
			β_x	−0.094	0.773	0.606		0.948	−0.026	0.504	0.255	
	30%	500	β_z	0.042	0.549	0.303	0.932	0.970	−0.516	0.268	0.338	0.748
			β_x	0.009	1.170	1.368		0.967	−0.005	0.730	0.533	
		1000	β_z	0.055	0.378	0.146	0.723	0.945	−0.517	0.190	0.304	0.653
			β_x	−0.027	0.857	0.736		0.946	−0.024	0.512	0.263	
continuousuniform	15%	500	β_z	−0.018	0.493	0.243	0.566	0.960	−0.508	0.254	0.323	0.611
			β_x	−0.082	0.586	0.351		0.973	0.001	0.395	0.156	
		1000	β_z	−0.024	0.359	0.130	0.433	0.959	−0.517	0.181	0.301	0.570
			β_x	−0.083	0.452	0.211		0.966	0.010	0.270	0.073	
	30%	500	β_z	0.064	0.523	0.277	0.642	0.961	−0.511	0.264	0.331	0.621
			β_x	−0.064	0.681	0.468		0.950	0.003	0.409	0.167	
		1000	β_z	0.027	0.371	0.138	0.494	0.955	−0.524	0.183	0.308	0.578
			β_x	−0.052	0.549	0.304		0.957	0.008	0.275	0.076	
discreteBernoulli	15%	500	β_z	−0.009	0.474	0.225	0.434	0.957	−0.514	0.253	0.328	0.569
			β_x	−0.024	0.349	0.122		0.971	0.013	0.233	0.054	
		1000	β_z	−0.013	0.351	0.124	0.333	0.937	−0.518	0.180	0.301	0.548
			β_x	−0.038	0.252	0.065		0.948	0.017	0.154	0.024	
	30%	500	β_z	0.017	0.503	0.253	0.436	0.957	−0.512	0.263	0.331	0.571
			β_x	0.002	0.347	0.120		0.957	0.015	0.243	0.059	
		1000	β_z	0.000	0.346	0.120	0.308	0.952	−0.522	0.181	0.306	0.553
			β_x	0.016	0.233	0.055		0.963	0.015	0.159	0.026	

References

1. Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**(434), 444–455 (1996)

2. Beran, R.: Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley (1981)

3. Chernozhukov, V., Hansen, C.: An IV model of quantile treatment effects. *Econometrica* **73**(1), 245–261 (2005)

4. Giacomini, R., Politis, D.N., White, H.: A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Economet. Theor.* **29**(3), 567–589 (2013)
5. Huang, X., Xiong, C., Jiang, T., Lu, J., Xu, J.: Efficient estimation and inference in the proportional odds model for survival data. *Mathematics* **10**(18), 3362 (2022)
6. Li, S., Peng, L.: Instrumental variable estimation of complier causal treatment effect with interval-censored data. *Biometrics* **79**(1), 253–263 (2023)
7. McLain, A., Ghosh, S.K.: Estimation of time transformation models with Bernstein polynomials. Technical report, North Carolina State University. Dept. of Statistics (2009)
8. Murphy, S., Rossini, A., van der Vaart, A.W.: Maximum likelihood estimation in the proportional odds model. *J. Am. Stat. Assoc.* **92**(439), 968–976 (1997)
9. Panaro, R.V.: *spsurv*: an R package for semi-parametric survival analysis. arXiv preprint [arXiv:2003.10548](https://arxiv.org/abs/2003.10548) (2020)
10. Tedesco, L., Beyhum, J., Van Keilegom, I.: Instrumental variable estimation of the proportional hazards model by presmoothing. arXiv preprint [arXiv:2309.02183](https://arxiv.org/abs/2309.02183) (2023)
11. Wüthrich, K.: A comparison of two quantile models with endogeneity. *J. Bus. Econ. Stat.* **38**(2), 443–456 (2020)
12. Yu, W., Chen, K., Sobel, M.E., Ying, Z.: Semiparametric transformation models for causal inference in time-to-event studies with all-or-nothing compliance. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **77**(2), 397–415 (2015)



Measuring Political Cleavages Through the Decomposition of Diversity

Giulio Giacomo Cantone¹ and Venera Tomaselli²(✉)

¹ Business School, University of Sussex, Brighton, UK
ggc23@sussex.ac.uk, prgcan@gmail.com

² Department of Economics and Business, University of Catania, Catania, Italy
venera.tomaselli@unict.it

Abstract. This study offers a method for the measurement of political cleavages by decomposing the measure of diversity in the opinions of survey respondents, in *within* and *between* parts across sub-populations formed through demographic categories. A synthetic population is generated by adopting as a seed the responses to a questionnaire from a non-representative sample. Two different kinds of cleavage are considered: the first measures how far apart two finite and logically adjacent sub-populations are in terms of the opinions surveyed through the questionnaire; the second measures how much a G categorical grouping variable is relevant to observing divisive opinions. In the conclusions, an extension of the method is offered for sample estimates.

Keywords: Diversity · decomposition · synthetic population · cleavages

1 Introduction

The concept of diversity is close to being universal across scientific disciplines [1]. In Statistics it is often associated with the measurement of the ‘qualitative variation’ [2] in the probability mass of categorical variables, i.e. the heterogeneity across the frequencies of nominal classes associated with the X analytical unit. Diversity is often measured through a measure close to the Herfindahl-Hirschman index, the complement of the rate of repeats [3, 4]:

$$\Delta(X) = 1 - \sum_i p_i(X)^2 \quad (1)$$

where i denotes the categories of X , so $p_i(X)$ is the relative frequency of the category i observed in X - hence this mention of the concept of ‘repetition’.

Across social sciences, the interest in diversity is ubiquitous. Alongside “Equity” and “Inclusion”, “Diversity” is a cornerstone of contemporary progressive Political Philosophy. Economic diversification (e.g. in portfolios or market

shares) is advocated to avoid the composite risks associated with the concentration of few actors in decision-making. In Political Science the interest is in the measurement of the so-called ‘effective’ number of parties in a Parliament [5].

Ideally, the measurement of diversity of opinions works as an indicator of the pluralism of ideas in a population. This study offers an approach to the measurement of political cleavages, which are divisions across citizens that define political factions with competing political agendas [7]. Indicators for the extension of the cleavages are reached by decomposing the measure of diversity in opinions, on a population of survey responses, in *within* and *between* parts across sub-populations formed through the demographic categories of the respondents.

2 Methods

Let’s assume a Q questionnaire with q questions that can be answered through an ordinal scale of m_q values, from 0 to m . Then exist $k = m^q$ possible ways to answer Q questionnaire. The set $I = \{i_1, i_2, \dots, i_k\}$ contains all the univocal combinations of answers that a respondent x of the population X can provide to the Q , i.e. it is the set of the k possible opinions expressed as unique $\mathbf{i}$ vector that answers Q . The formalism $m_q(i)$ then symbolises the value of the position q in the vector of answers $\mathbf{i}$ forming the opinion i .

Once all the elements of X respond to Q , for each opinion i it can be derived its associated frequency p_i . The straightforward computation of the rate of repeats (Eq. 1) would ignore that the conceptual d -distance between opinions is not uniform since some answers are closer in the ordinal scale. Stirling [8] solves this issue reformulating Eq. 1 as follows:

$$\Delta(X | D) = \sum_{(i,j)} p_i \cdot p_j \cdot D(i, j) \quad (2)$$

where $j \in J : J = I$, and D , called ‘factor of disparity’ is the square matrix of distances D , such that $\arg \max d(i, j) = 1$ and $\arg \min d(i, j) = 0$. This minimum of 0 is necessarily reached in the diagonal when $i = j$.

Given how the questionnaire is structured, a formulation of $d(i, j)$ is derivable from the normalised absolute vector differences of opinions [6]:

$$d(i, j) = \frac{\sum_q |\mathbf{i} - \mathbf{j}|}{m \cdot q} \quad (3)$$

where the normalising denominator satisfies the requirements of Stirling formula.

2.1 Decomposition of Diversity Across Sub-populations

Two different kinds of cleavage are considered. The first is specific *group cleavages* that measure how finite fractions of a population (groups) that are logically adjacent (for example, they share the trait of being ‘young’ or of being

‘cosmopolitans’) are distant in terms of the opinions surveyed through the questionnaire. In this case, the cleavage is always between two groups. The second kind is *trait cleavages* regarding how divided the opinions are along a whole G categorical grouping variable [2, 6].

The two analyses converge to a unique operation in the simplest case of a population divided between two X_g groups by a G binary classification. In this case, exists two intra-group diversities $\Delta(X_g)$ for each X_g , and a singular inter-group diversity that is equal to their average $\bar{\Delta}(X | G)$. While it can happen than $\exists \Delta(X_g) : \Delta(X_g) > \Delta(X)$, it holds that $\bar{\Delta}(X | G) < \Delta(X); \forall X$. By analogy with the canonical decomposition of variance across groups [2], the average $\bar{\Delta}(X | G)$ is called *within* Diversity; from this quantity, the *between* component of the diversity of X is derived as:

$$\Delta_G(X) = \Delta(X) - \bar{\Delta}(X | G) \quad (4)$$

In this study, the ratio:

$$\delta(G | X) = \frac{\Delta_G(X)}{\Delta(X)} \quad (5)$$

is considered the measure of the *trait cleavage* within the population X , due to the G grouping factor, and it is also the only *group cleavage* being only two groups across a binary variable.

For n_g states of the G variables, there are $\frac{n \cdot (n-1)}{2}$ specific *cleavages*, which are still derived as the quota of the *between* diversity of coupled states (e.g. groups) over the diversity of their union; in the extended form:

$$\delta(X_g, X_{g'}) = \left[\Delta(X_g \cup X_{g'}) - \frac{\Delta(X_g) + \Delta(X_{g'})}{2} \right] \cdot \frac{1}{\Delta(X_g \cup X_{g'})} \quad (6)$$

This derivation can be adopted for $\mathbf{G} : \{G_1, G_2, \dots, G_n\}$ multivariate grouping factors, where a X_g group is the intersection of singular states of $\mathbf{G}$, for example, the group of ‘young ladies’ when the grouping factors are gender and age. In this case, mathematically exist *group cleavages* for any couple $(X_g, X_{g'})$, nevertheless given the multivariate structure of the grouping structure, the most sensible comparison is among all the cleavage of a group with its adjacent groups, i.e. groups sharing at least a state in common.

3 Application

A synthetic population is generated by adopting as a seed the responses to a questionnaire from a non-representative sample ($n = 465$) of Italian citizens collected in the week after the National Elections, 22th September 2022. Among the answers collected, respondents reported anonymously the party they voted for the election of the Lower Chamber, and they proceeded to score on a scale in $(0 : 3)$ their opinion, on 5 q -questions (ergo $k = 4^5 = 1024$):

1. Is immigration a net positive event for Italy?
2. Are you satisfied with how Italy contrasted the COVID-19 pandemic?
3. Do you support basic income (*Reddito di Cittadinanza*) as a welfare policy?
4. Should Italy support Ukraine in the conflict against Russia?
5. Do you support stricter laws (“Legge Zan”) against homophobia?

The reported votes have been categorised into four political g blocks. Next to the name of the block is reported the percentage of votes for each block computed on votes validly expressed for the Lower Chamber:

- Conservatives (44%): Fratelli d’Italia; Lega per Salvini Premier; Forza Italia; Noi Moderati.
- Social-Democrats (24%): Partito Democratico - Italia Democratica e Progressista; Alleanza Verdi e Sinistra; Impegno Civico - Centro Democratico; Südtiroler Volkspartei; Noi di Centro.
- Skeptics (21%): Movimento Cinque Stelle; Italexit, Unione Popolare; Italia Sovrana e Popolare; Vita; Partito Comunista Italiano.
- Liberals (11%): Azione - Italia Viva; +Europa; Sud Chiama Nord.

The generating process of the synthetic population is the following: a random g is drawn with a probability equal to the reported percentages of its votes. Within this block, an a respondent is drawn with uniform probability. Then, for each q question, a random number is drawn with the probability function of a Binomial model of $m = 3$ trials and a probability of a successful trial equal to $\frac{\bar{m}_q(g) + m_q(a)}{6}$, where $\bar{m}_q(g)$ is the average score to q within the g block. The five randomly generated numbers are the values of the new synthetic respondent a' , who is also assigned to the randomly drawn block g . This operation is repeated for $N := k = 1024$ times to generate the full synthetic population in the 4 blocks.

3.1 Results

The political agenda of these four blocks can be situated across two binary grouping variables: Geopolitics and Economic Doctrine.

- Conservatives are localists (nationalists) and they favour free markets.
- Social-Democrats are cosmopolitan and they favour planned economies over free markets.
- Skeptics share a distrust for globalism and preference for localism and also a preference for robust welfare over free markets.
- Liberals are cosmopolitans and they favour free markets over planned economies.

These considerations have been formalised in a rhomboidal scheme in Fig. 1.

It can be noticed how the *trait cleavages* are relatively uninformative: 13.5% and 16% of Diversity of opinion along Geopolitics and Economic doctrine are low values, meaning that these two dimensions alone do not explain very well the extent of diversity of opinions surveyed by the five questions. In other words,

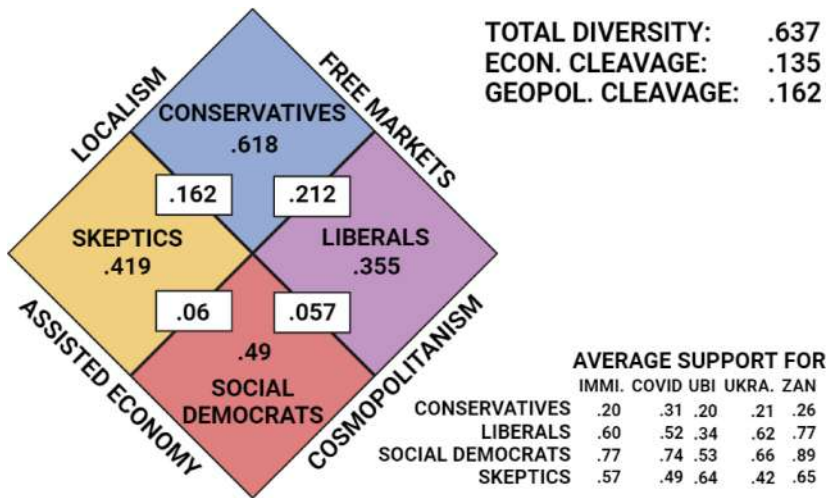


Fig. 1. Statistics on Diversity, Cleavages, and Opinions across the groups. The number close to the label is the value of Rao-Stirling Diversity [8] within the group, while the values in the white boxes on the borders are the measured cleavages. The D matrix is derived through the method in Eq. 3. The diversity values are predicted by the size of the group. This is an effect of diversity increasing stochastically with the probability of an opinion i being non-zero frequent in the population. Nevertheless, the ratio between the size and the diversity has no canonical statistical interpretation. ‘UBI’ stands for the question about basic income, ‘Zan’ about homophobia. The values for the support for the opinions are the group averages normalised by $m = 3$.

the variance in the provided answers transcends conventional political blocks. In general, low trait cleavages would predict political fluidity in the electors because it is a sign of low ideological divisions. This is consistent with large swings in electoral preferences and political alignments after 2013 in Italy.

Nevertheless, *group cleavages* are more useful: the opinions of the Conservatives are separated from those of Skeptics and Liberals, compared to how close the opinions of these with Social-Democrats are. It can be derived that in the population the general cleavage is between the large group of Conservatives and all the other minority blocks. This observation is consistent with the average support for single opinions, where Conservatives show average support for the issue inferior to .5, while other blocks are almost always superior. Hence, *group cleavages* act as a summarisation of multifaceted differences of opinion.

4 Future Development and Limitations

The proposed methods measure exact values for cleavages on a fully covered population through questionnaires. Nevertheless, this is never feasible and only sample statistics can be derived [6]. Thus, the exact determination of confidence

intervals for group cleavages is complicated by the stratified nature of the population [9].

The application has been conducted to showcase intuitive Monte Carlo methods to derive representative sample statistics for a surveyed population. For simplicity, the application assumes a synthetic population of a size equal to the number of opinions, that is of many magnitudes inferior to the amount of valid votes at the Italian Election. Instead for a representative study, known N of the target population, many synthetic populations of $\sim N$ size can be simulated from a representative sample seed with differentiated methods for generating the synthetic variance of answers. Then, the median values across the synthetic populations can be assumed as representative of the cleavages and the confidence intervals can be derived likewise at any percentile.

The application ignores by design the quota of people who have abstained. Let's be these symbolised as X_0 . Then for each group in X_g exists a technical cleavage with X_0 . Since the cleavage $d(X_0, X_g)$ would measure in this case the political distance between abstained and political factions, it could be worth testing if given a Q set of questions, the cleavages are predictive of future swing of the vote of abstained into a pre-existing g -block.

Finally, the method could incur issues of computability and bias due to the inherent complexity of eliciting unique opinions for each i unique vector of answers. k raises exponentially with the number of questions, and generally questionnaires on opinions range between eight and twenty-five questions. The formula for the matrix of distances in Eq. 3 deals with ordinal scales, but it does not account for the factor of disparity of the correlations among the answers to the questions.

The application assumed the five questions as five statistically independent dimensions of the political opinion of the respondents but in reality, these dimensions are never strictly independent. For example, the support for immigration is uniformly correlated with the support for stricter laws against homophobia, even if the two dimensions are sufficiently well distinct, and as a consequence, they may be seen as independent. A common solution to either issue would be a selection of a few, low-correlated questions.

References

1. McDonald, D.G., Dimmick, J.: The conceptualization and measurement of diversity. *Commun. Res.* **30**(1), 60–79 (2003). <https://doi.org/10.1177/0093650202239026>
2. Teachman, J.D.: Analysis of population diversity: measures of qualitative variation. *Sociol. Method. Res.* **8**(3), 341–362 (1980). <https://doi.org/10.1177/004912418000800305>
3. Patil, G.P., Taillie, C.: Diversity as a concept and its measurement. *J. Am. Stat. Assoc.* **77**(379), 548–561 (1982). <https://doi.org/10.1080/01621459.1982.10477845>
4. Rousseau, R.: The repeat rate: from Hirschman to Stirling. *Scientometrics* **116**(1), 645–653 (2018). <https://doi.org/10.1007/s11192-018-2724-8>
5. Hanretty, C.: Party system polarization and the effective number of parties. *Elect. Stud.* **76**, 102459 (2022). <https://doi.org/10.1016/j.electstud.2022.102459>

6. Lieberman, S.: Measuring population diversity. *Am. Sociol. Rev.* **34**(6), 850–862 (1969). <https://doi.org/10.2307/2095977>
7. Stoll, H.: Social cleavages and the number of parties: how the measures you choose affect the answers you get. *Comp. Polit. Stud.* **41**(11), 1439–1465 (2008). <https://doi.org/10.1177/0010414007305813>
8. Stirling, A.: A general framework for analysing diversity in science, technology and society. *J. R. Soc. Interface* **4**, 707–719 (2007). <https://doi.org/10.1098/rsif.2007.0213>
9. Caso, C., Gil, M.A.: The Gini-Simpson index of diversity: estimation in the stratified sampling. *Commun. Stat. Theory Methods* **17**(9), 2981–2995 (1988). <https://doi.org/10.1080/03610928808829784>



The Ripple Effect: How Pre-surgery Information Shapes Patient Satisfaction in Breast Cancer Surgery

Costanza Tortú^{1(✉)}, Chiara Seghieri¹, and Francesca Ferré²

¹ Department L'EMbeDS, Institute of Management, Sant'Anna School for Advanced Studies, Pisa, Italy

`costanza.tortu@santannapisa.it`

² Department of Biomedical Sciences for Health, University of Milan, Milan, Italy

Abstract. Breast cancer poses a global burden, with thousands of women undergoing surgery annually. Recently, there's a growing acknowledgment of the importance of developing a patient-centered care, leading to increased interest in Patient-Reported Outcome Measures (PROMs): this, in the oncology setting, translates into the distribution of validated tools like the BREAST-Q survey, that is specifically designed for breast cancer patients. Using data collected through the BREAST-Q survey in the University Hospital of Pisa (Italy) between 2018 and 2022, this research explores the causal relationship between the quality of pre-operative information provided by the surgeon to the patient and the post-surgery patient satisfaction with the results after surgery. We employ Item Response Theory (IRT) models to estimate patient-level fulfillment with pre-surgery information and post-surgery overall satisfaction. Then, using estimated traits, we assess the causal effect of information quality on post-surgery breast satisfaction through an Exposure Response Curve (ERC), accounting for various individual pre-exposure characteristics.

Keywords: causal inference · breast cancer · item response theory · exposure response curve

1 Introduction

Breast cancer is a global health concern, with thousands of women undergoing breast cancer surgery annually worldwide. In recent years, there has been a growing recognition of the necessity to develop patient-centered care and this translates in an increased interest in Patient-Reported Outcome Measures (PROMs) [1]. In the context of breast cancer surgeries, the significance of collecting quality-of-life outcomes is underscored by the existence of various internationally validated surveys, such as BREAST-Q [2].

In this work, we estimate the causal relationship between the quality of pre-surgery information provided by the surgeon and the patients' breast satisfaction

after surgery (it is a clinical outcome, woman after surgery directly report their perception of the appearance of their breast). Both these aspects are explored in the BREAST-Q survey, where women are asked to provide their answers to series of items, measuring these two aspects. By analyzing responses to these items, we measure two quantities at the patient-level representing their (latent) overall level of cheerfulness with pre-surgery information and their post-surgery breast satisfaction: both these quantities are estimated by using a purely data-driven approach based on Item Response Theory (IRT) models [3], specifically tailored to ordinal items - known as Graded Response Models (GRMs) [4] -.

Once estimated these two latent traits, we estimate the causal effect of the quality of information on post surgery satisfaction through a Exposure Response Curve (ERC) [5], while controlling for selected individual characteristics (socio-demographic characteristics, information on the specific diagnosis and surgery, characteristics of the pre-surgery clinical profile).

2 Data

Data include individual responses to the BREAST-Q survey, collected between 2018 and 2022 in the University Hospital of Pisa (AOUP). Women that are eligible for participation in the survey are those admitted for inpatient or day hospital care with a: i) primary diagnosis of invasive malignant breast tumor (ICD-9-CM 174*) undergoing surgical intervention for mastectomy (ICD-9-CM 85.4*) combined with reconstruction procedures or tissue expander insertion (ICD-9-CM 85.33, 85.35, 85.7, 85.85, 85.95) or implantation of prosthesis (85.53, 85.54) performed during the hospitalization; ii) primary diagnosis of invasive or in situ malignant breast tumor (ICD-9-CM 174*, 233.0) undergoing surgical intervention for quadrantectomy (ICD-9-CM 85.20, 85.21, 85.22, 85.23, 85.24, 85.25) or reduction mammoplasty (85.31, 85.32). Eligible participants include women with primary tumors, recurrence, or re-intervention for residual disease, positive lymph nodes, aesthetic improvement, and reconstructions for prophylaxis. The final sample consists in 482 women, who meet the eligibility criteria and who complete both a pre-surgery survey and a post-surgery survey. Data provide information i) on the rates assigned by women to the statements related to the quality of pre-operative information provided by the surgery and to the level of post-surgery breast satisfaction; ii) on some demographic characteristics and iii) on some clinical aspects related to the diagnosis, to the surgery, to the clinical picture of the patient and to her general health status. All items related to i) and ii) are graded from 1 to 4 (1: Totally unsatisfied; 2: Partially unsatisfied; 3: Partially satisfied; 4: Totally satisfied).

3 Methods

3.1 Set Up

The sample $\mathcal{N}$ includes N patients, where for a given patient j with $(j : 1, \dots, N)$ we observe i) her assigned rates to the items in $\mathcal{I}_1$, where the generic item is

indexed as i_1 , with $(i_1 : 1, \dots, I_2)$ and is designed to evaluate her perceived quality of the pre-operative information received by the surgeon, ii) her assigned rates to the items in $\mathcal{I}_2$, where the generic item is indexed as i_2 , with $(i_2 : 1, \dots, I_2)$ and is designed to evaluate her post-operative breast satisfaction, iii) a P -dimensional vector of baseline characteristics $\mathbf{X}_j = \{x_{j1}, \dots, x_{jP}\}$ including demographics and clinical features. All items in $\mathcal{I}_1$ and $\mathcal{I}_2$ provide $K = 4$ response categories.

3.2 Item Response Theory

To forecast an individual's latent perception of the quality of information received from the doctor and breast satisfaction after surgery, we implement two separate IRTs on the individual responses to the items $\mathcal{I}_1$ and $\mathcal{I}_2$. IRTs are commonly used in various social science fields to estimate latent traits from individual responses to a series of items. These items can be binary, categorical, or ordinal, with corresponding adaptations of IRT models. Ordinal items are modelled according to a GRM. In a GRM, each item is represented by its discrimination parameter and specific thresholds, which define difficulty parameters separating ordered response categories. Formally, a GRM involves I items $(i : 1, \dots, I)$, each featuring K response categories $(k : 1, \dots, K)$, and models the probability of observing a response R greater than or equal to k for item i and individual j ($j : 1, \dots, N$). This probability is expressed as:

$$P(R_{ij} \geq k | \theta_j) = \frac{\exp(a_i(\theta_j - b_{ik}))}{1 + \exp(a_i(\theta_j - b_{ik}))}$$

where a_i represents the discrimination parameter of item i , b_{ik} is the k -th threshold for item i , θ_j is the value of the latent trait for person j . The threshold b_{ik} can be interpreted as the difficulty in responding to a category equal to or greater than k for item i . Meanwhile, the a_i parameters provide insights into the correlation between the latent trait and the likelihood of responding to a category higher than k . In this setting, we estimate two separate GRMs to predict the latent perception on the quality of pre-operative information provided by the surgeon and the latent post-surgery breast satisfaction: we denote these predicted traits as $\hat{\theta}_{\mathbf{I}_1}$ and $\hat{\theta}_{\mathbf{I}_3}$, respectively. Note that, given the properties of GRMs, the predicted latent traits $\hat{\theta}_{\mathbf{I}_1}$ and $\hat{\theta}_{\mathbf{I}_2}$ both follow a Standard Normal Distribution within the population - i.e. $\theta \sim \mathcal{N}(0, 1)$ -.

3.3 Causal Inference for Continuous Exposures

Once obtained the predicted latent traits we estimate the causal effect of the quality of pre-operative information on post-operative satisfaction, where the exposure and the outcome are represented by the predicted traits $\hat{\theta}_{\mathbf{I}_1}$ and $\hat{\theta}_{\mathbf{I}_2}$, respectively. This setting implies a continuous exposure (the quality of pre-operative information), in an observational setting, where data also provide full information on a covariate matrix $\mathbf{X}$. We outline the causal estimand of interest

under the Potential Outcomes (POs) framework, where the continuous intervention W_j denotes the quantity $\hat{\theta}_{\mathbf{I}_{11}}$ (with $W_j \in \mathbb{R}^+$) and the outcome Y_j identifies $\hat{\theta}_{\mathbf{I}_{12}}$. In the POs setting, for each unit j in $\mathcal{N}$ we can identify a random variable $Y_j(w)$ representing the potential outcome of unit j when exposed to a given level $w \in \mathcal{W}$ of the exposure, where the set $\mathcal{W}$ represents the domain of the exposure. The fundamental problem, which is often known as the *fundamental problem of causal inference*, is that the potential outcome $Y_j(w)$ is missing for all $w \neq W_j$. Since we are not able to directly estimate $Y_j(w)$ as we only observe $Y_j(W_j)$ we have to shift our focus on estimating the population Exposure-Response Curve (ERC) $R(w) = \mathbb{E}[Y_j(w)]$. There exist various strategies to estimate ERCs, here we follow the approach proposed by [6], who develop a nonparametric doubly-robust estimator of the exposure response curve using the pseudo-outcomes fit.

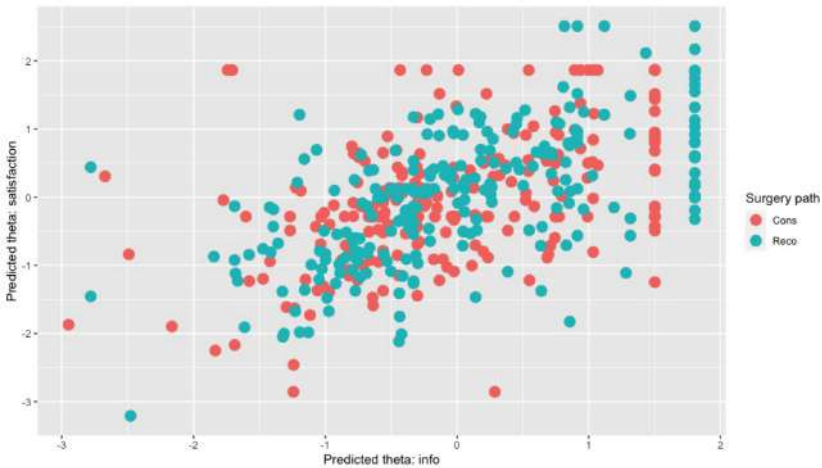


Fig. 1. Scatterplot, joint distribution of predicted thetas: the x-axis represents the predicted quality of pre-operative information, while the y-axis represents the predicted post-surgery satisfaction. Colors refer to the broader class of breast cancer surgeries they refer to (reconstructive or conservative)

4 Preliminary Results

Preliminary results indicate that i) the implemented Graded Response Models (GRMs) accurately predict both the latent traits measuring the quality of pre-operative information provided by the surgeon and the overall post-surgery breast satisfaction; ii) there is a strong positive association between the two estimated latent traits (see Fig. 1); iii) these findings hold for both surgical paths tailored for breast cancer patients (conservative and reconstructive surgery). iv) the estimated ERC suggests that higher quality pre-operative information leads

to an increase in overall post-surgery breast satisfaction, with the estimates being particularly accurate in the central part of the exposure domain.

5 Conclusions

In the last few years, PROMs data specific to breast surgery have gained significant popularity and validity. They have been increasingly used as information on perceived health status and quality of life after surgery, complementary to the traditional clinical outcomes such as readmissions and mortality. Findings from this study demonstrates a causal effect of effective communication between the surgeon and the patient before the surgery with increased patient reported post-operative satisfaction. Nonetheless, in a time in which healthcare systems have to become more efficient to face the increased demand of care on one side with limited resources on the other, physicians might be pressured to maximize the number of patients they see thus decreasing the time spent with them. In this context, the study sheds light on the importance of physician-patient communication as a crucial factor in patient reported outcomes and present evidence-based recommendations to healthcare managers and policy makers on strategies and efforts to maximize the time physicians dedicate to build the relationship with their patients and, at the same time, to improve their interpersonal and communication skills. To the best of our knowledge this is the first study that analyze the causal affect of the satisfaction with the information received and patients' reported outcomes following breast surgery.

References

1. Meadows, K.A.: Patient-reported outcome measures: an overview. *Br. J. Community Nurs.* **16**(3), 146–151 (2011)
2. Pusic, A.L., Klassen, A.F., Scott, A.M., Klok, J.A., Cordeiro, P.G., Cano, S.J.: Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. *Plast. Reconstr. Surg.* **124**(2), 345–353 (2009)
3. Baker, F.B.: *The Basics of Item Response Theory*, ERIC (2001)
4. Samejima, F.: The general graded response model. In: *Handbook of Polytomous Item Response Theory Models*, pp. 87–118. Routledge (2011)
5. Wang, B., et al.: The impact of long-term PM 2.5 exposure on specific causes of death: exposure-response curves and effect modification among 53 million US Medicare beneficiaries. *Environ. Health* **19**, 1–12 (2020)
6. Cork, M., Mork, D., Dominici, F.: Methods for estimating the exposure-response curve to inform the new safety standards for fine particulate matter. *arXiv preprint [arXiv:2306.03011](https://arxiv.org/abs/2306.03011)* (2023)



Employment Trajectories and Fertility in Italy

Marco Tosi^(✉), Fausta Ongaro, and Silvia Meggiolaro

Department of Statistical Sciences, University of Padova, Padua, Italy

marco.tosi@unipd.it

Abstract. This study investigates whether employment/education histories by age 30 are associated with the cumulated fertility at age 30 and with the completed fertility at age 45. We use data from the ISTAT survey “Families and Social Subjects” and adopted a sequence analysis approach to describe the employment/education histories of individuals aged 45–60, while accounting for complex time-related interdependencies among employment episodes. We compare different sequences by using optimal matching methods and perform cluster analysis to identify the most frequent typologies of employment/education trajectories. Preliminary results from Poisson regression models suggest that men and women entering the labor market at a later age have, on average, fewer children by age 30 compared to those with other employment/education trajectories. However, this negative association disappears when we analyze fertility at age 45, suggesting that individuals who postpone their entrance into the labor market are able to recover in the long run. The fertility behaviors of individuals with uncertain employment histories do not differ from those of people with more stable employment histories. Lastly, women who never worked tend to be characterized by larger families at age 30 but not at age 45. This study shows the importance of analyzing the complex interplay of employment episodes in early adulthood to predict completed fertility later in life.

Keywords: Fertility · Employment History · Sequence Analysis · Italy

1 Introduction

Many studies have investigated the relationship between employment and fertility, with the assumption that the lack of income or the economic uncertainty related to disadvantaged occupations are crucial constraints driving fertility decisions in industrialized countries. However, empirical results are mixed. While the macro-perspective suggests that a fertility decline may be associated with periods of economic hardship (see, for example, Billari and Kohler, 2004), the micro-evidence presents more mixed results (Cazzola et al. 2016). Studies focusing on unemployment report no effect on the reproductive behaviors of women (Kreyenfeld 2010), while others show a detrimental effect of job loss and atypical employment on the transitions to motherhood (Barbieri et al. 2012); a positive impact of unemployment among lower educated women is found as well (Kreyenfeld, 2010). Similarly, results for men are contradictory too (Özcan et al. 2010; Vignoli et al. 2012).

However, most previous studies focus on individual work statuses, as static measures, which is a limitation given that perceptions of economic vulnerability or instability are likely be influenced not only by current situation but also by past employment experience as well as by future scenarios or prospects. Moreover, such studies mainly address the single transitions from a parity to another – often considering the transition to the first child – in which the timing and quantum effects are mixed. Only recently, indeed, researchers have studied the role of employment histories on completed fertility, but, again, mixed results emerge (see, for example, Ciganda 2015).

The current paper is inserted in this line of research, with the aim of analyzing the relationship between individual employment/education history and completed fertility for individuals born between 1956 and 1971. More specifically, we are interested in examining whether the characteristics of the employment/education biography till the age of 30 is associated with the cumulated fertility at the age of 30 and the completed fertility (at age 45).

2 Data and Methods

The current study uses data from the survey “Families and Social Subjects”, conducted by the Italian National Institute of Statistics (ISTAT) in 2016. The survey is based on a representative sample at the national level of 24,753 people over the age of 18. Besides information on reproductive history of the interviewees, the survey provides retrospective data on each employment episode, from the first job to the eleventh one, allowing the construction of the complete employment histories of individuals. In this paper, we analyze 3525 men and 3700 women aged 45–60 at the time of interview born between 1956 and 1971.

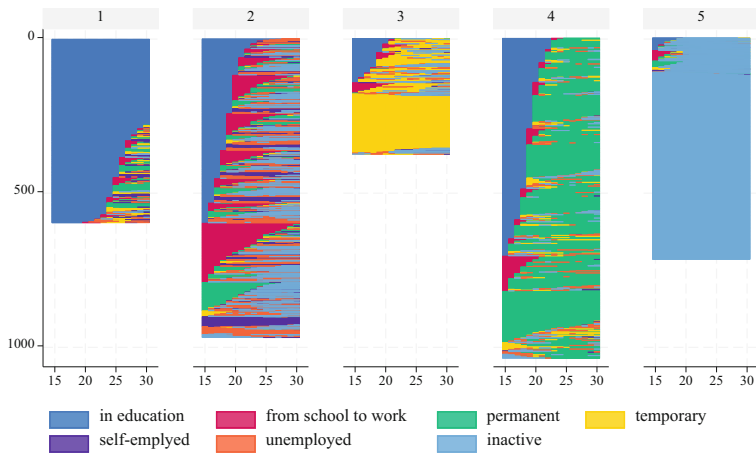
Sequence analysis techniques were applied to describe employment trajectories from age 15 to 30, a life stage in which youths typically establish themselves in the labor market and experience the transition to adulthood. Our approach consists in distinguishing between 7 life-course states: in education (based on the year of completed education, plus additional years spent in the educational system); from school to work (the time spent between leaving education and starting the first job); permanent job; temporary job (or fixed-term contracts); self-employment; inactivity (leaving a job for family reasons or never worked); and unemployment. We analyze these sequences by using Optimal Matching to compare each sequence with the others and calculate a matrix of distances and similarities between sequences. The employment/education trajectories were, then, clustered in groups on the basis of their distances and similarities. The derived typology was used as an independent variable to predict fertility at age 30 and fertility at age 45 in Poisson regression models. Control variables included in the models are: number of siblings (0–4); parental education (up to lower secondary; secondary; tertiary); mother’s and father’s employment status at age 14; cohort (1956–61; 1962–66; 1967–71), Italian macro-region (North; Centre; South), and marital status (i.e. being married).

3 Results

3.1 Employment Histories Groups

Figure 1 describes the clusters obtained for women's and men's employment histories. Among women, five clusters were identified: G1 '*Late entry into the labor market*' (16.1%) is characterized by enrolment in higher education, and a postponement of paid work (47% of them are in education at age 30); G2 '*Early exit from the labor market*' (26.2%) is characterized by long periods of inactivity before entering the labor market and an early exit from the labor market; G3 '*Uncertainty*' (10%) is a group defined by fixed-term contracts and some interruptions of inactivity and unemployment; G4

Women



Men

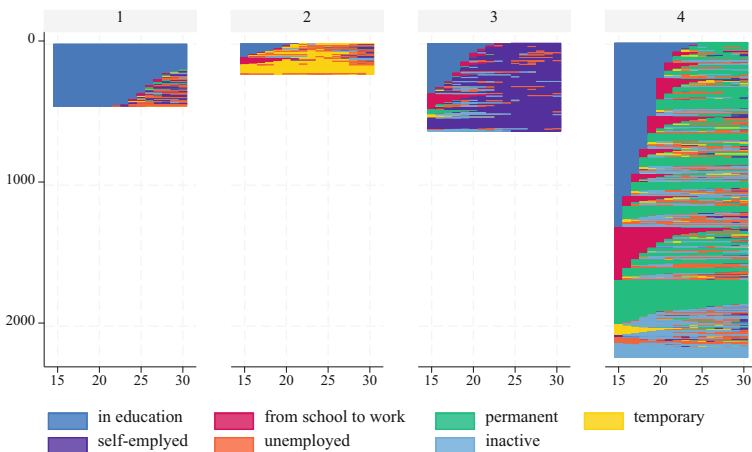


Fig. 1. Clusters of employment careers.

‘*Stable employment*’ (28%) is characterized by permanent contracts; G5 ‘*Inactivity*’ (19.4%) is a group of women never entered in the labor market. Among men, four clusters are identified: G1 ‘*Late entry into the labor market*’ (12.5%) is characterized by enrolment in higher education and a delayed entry into employment; G2 ‘*Uncertainty*’ (6.1%) is defined by individuals with fixed-term contracts and some interruptions due to unemployment; G3 ‘*stable self-employment*’ (17.7%) are self-employed men (with also some episodes of unemployment); and finally, G4 ‘*Stable employment*’ (63.7%) are men in stable employment (with also some episodes of inactivity, unemployment and temporary employment at the very beginning).

3.2 Employment Histories and Fertility

Poisson regression models were estimated on the parity achieved by individuals at age 30 and 45. The results (Fig. 2) show that the type of employment/education history is associated with individual fertility at age 30. Both men and women belonging to the group “*late entry*” into the labour market have, on average, fewer children by age 30 than those of the other groups. This negative association becomes smaller when we control for being married at age 30 and the impact is stronger for men. Moreover, the negative association found at age 30 disappears when we analyse fertility at age 45, thus suggesting that individuals who postpone their entrance into the labor market are able to recover in the long run. The cluster “*uncertainty*” is not associated with lower fertility levels, as one can expect; instead, it seems to foster women’s fertility in the first part of life (compared to the category “*stable employment*”). Women who never worked (“*inactivity*”) tend to be characterised by larger families at age 30 but not at age 45.

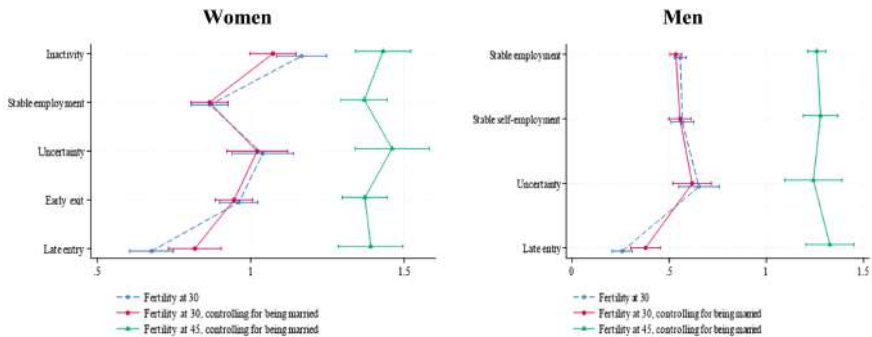


Fig. 2. Fertility at age 30 and age 45 estimated from Poisson regression models

4 Concluding Remarks

This study provides some preliminary evidence that the postponement of labor market entry has a short-term negative effect on fertility. These associations are partly mediated by marriage and tend to disappear when analyzing cumulated fertility at 45.

References

- Barbieri, P., Cutuli, G., Tosi, M.: Famiglie, mercato del lavoro e rischi sociali. Nascita di un figlio e rischi di transizione alla povertà tra le famiglie italiane. *Stato e mercato* **32**(3), 391–428 (2012)
- Billari, F.C., Kohler, H.-P.: Patterns of low and lowest-low fertility in Europe. *Popul. Stud.* **58**(2), 161–176 (2004)
- Cazzola, A., Pasquini, L., Angeli, A.: The relationship between unemployment and fertility in Italy: a time-series analysis. *Demogr. Res.* **34**(1), 1–38 (2016)
- Ciganda, D.: Unstable work histories and fertility in France: an adaptation of sequence complexity measures to employment trajectories. *Demogr. Res.* **32**, 843–876 (2015)
- Kreyenfeld, M.: Uncertainties in female employment careers and the postponement of parenthood in Germany. *Eur. Sociol. Rev.* **26**(3), 351–366 (2010)
- Özcan, B., Mayer, K.U., Luedicke, J.: The impact of unemployment on the transition to parenthood. *Demogr. Res.* **23**(29), 807–846 (2010)
- Vignoli, D., Drefahl, S., De Santis, G.: Whose job instability affects the likelihood of becoming a parent in Italy? A tale of two partners. *Demograph. Res.* **26**(2), 41–62 (2012)



Addressing Misclassification in the Assessment of Self-rated Health: Insights from the Elderly Population Italy

Eleonora Trappolini[✉], Maria Felice Arezzo, and Giuseppina Guagnano^{ib}

MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9,
00161 Rome, Italy

eleonora.trappolini@uniroma1.it

Abstract. Population health evaluation is vital, especially in aging populations. Self-rated health (SRH) is a widely-used tool in epidemiology, predicting health outcomes. However, concerns about its subjectivity and misclassification have been raised, particularly among disadvantaged groups. Measurement reliability is crucial for valid analysis, especially in health disparities research. This study proposes an approach to adjust regression estimates related to SRH and assesses the relationship between SRH and predictors among the elderly in Italy.

We use 2019 Italian data from the European Health Interview Survey for individuals aged 50 and above, and we conduct two separate analyses. We first model the association between SRH and a set of typical predictors by applying a standard ordered logit models, and then an adjusted one to account for potential misclassification in SRH and we compare the results.

In particular, we expect the association between the outcome variable and the independent variables will be weakened in the naive model when misclassification is present. In addition, higher education and stable relationship are expected to be associated with better SRH. This study addresses the subjectivity and potential misclassification in SRH. By applying adjusted models, we aim to obtain more accurate estimates in understanding SRH and its relationship with various predictors, contributing to better population health assessment.

Keywords: Misclassification · Self-rated health · Ordered logit · Elderly population · Italy

1 Introduction and Background

Evaluating population's health conditions has always been a crucial topic in demographic research. Recently, the interest towards population health was renewed as a consequence of the ageing process, which characterises most high-income countries.

In survey research on health there are a number of health indices to measure individuals' health status [1]. An alternative approach involves inviting survey participants to answer to a seemingly straightforward question 'How is your health in general?', in which respondents are asked to self-assess their health as 'very good,' 'good,' 'fair,'

‘bad,’ ‘very bad.’ This type of health measure, known as self-rated health (SRH), is a multifaceted tool commonly used for the evaluation of individuals’ health and it is also widely adopted due to its recognised substantial predictive and concurrent validity in the field of epidemiology and population health research. This validity is demonstrated by strong correlations with future health outcomes, including mortality and a range of morbidity, disability, and healthcare service utilisation indicators [2, 3]. Despite its extensive use in assessing health and its ability to predict adverse health outcomes, SRH as an indicator of individuals’ true health status is met with skepticism. This skepticism stems from the broader debate on the contrast between subjective perceptions and objective observations. Research suggests that self-reports of health and illness can be influenced by social experiences, potentially leading to inaccuracies, particularly among socially disadvantaged individuals from low- and middle-income backgrounds [4].

Measurement reliability is crucial for ensuring validity. Insufficient reliability corresponds to measurement errors, which potentially bias analytical results, especially when the construct serves as an independent or dependent variable in a regression model. In explaining an unreliable measure of SRH through an ordered regression model, it is known from the statistical literature that the estimators of the regression coefficients are inconsistent [5]. Therefore, to have consistent estimates of the regression parameters it is of paramount importance to rely upon an appropriate statistical methodology.

Moreover, the reliability levels may vary across different population subgroups, potentially affecting empirical comparisons. This is particularly significant when assessing health disparities, given the frequent use of SRH as a measure [6].

SRH, as a subjective health measure, distinguishes itself from most other health indicators due to the absence of formal, agreed-upon rules or definitions guiding the process of how individuals rate their health. As [3] suggests, SRH evaluation inherently requires the processing of information, the interpretation of meanings, and a process of selection. Despite being an individual, subjective process, it operates within the confines of a specific social and cultural context and draws upon the conceptual resources and patterns of representation provided by that environment.

In [7] work, an attempt was made to distinguish the impact of ‘true health’ from the impact of response styles in cross-country variations. This was achieved by using other health metrics present in the study to ‘adjust’ the self-ratings. However, this intriguing approach has its limitations as it fails to account for the unmeasured aspects of health not covered by ‘objective’ indicators and the intricacies of the evaluation processes [3].

There are also other studies that have explored the reliability of SRH [1, 8]. These studies revealed notable variations in SRH ratings when the health status questionnaire was administered on two separate occasions. However, it is important to acknowledge that perceptions and reporting of subjective health can differ significantly across countries [9]. In another study [6], the authors assessed the test-retest reliability of general SRH within a nationally representative sample of U.S. adults. They collected responses from two interviews conducted approximately one month apart and found that nearly 40% of respondents changed their health rating, indicating moderate test-retest reliability of SRH. They also examined reliability across key sociodemographic variables, including age, gender, race, and educational attainment.

Furthermore, other studies demonstrate the reliability of this health measure, suggesting that in the health status ratings, it is common to find respondents clustered in middle-range options rather than in extreme ones [10]. For instance, participants prefer ‘good’ over ‘very good’ to potentially conceal some reservations about their health status [11], or they may prefer ‘good’ over ‘fair’ as the latter is perceived as too negative [12].

Drawing upon this literature, the aim of this study is to propose an approach for adjusting the estimates of the regression coefficients and, consequently, the estimates of the misclassification probabilities associated with each SRH level. The approach will be employed to analyse the relationship between SRH and a set of typical predictors (e.g., educational level, marital status) among the elderly in Italy.

2 Data

For our study, we used the 2019 Italian data from the European Health Interview Survey (IT-EHIS, wave 3) carried out by the National Institute of Statistics (Istat). The IT-EHIS survey collects information on the health status, healthcare provision, health determinants and socioeconomic condition of the Italian population. The sample includes data on nearly 46,000 individuals aged 15 and above, residing in private dwellings. Given our study’s focus on the health of the elderly population, we restricted our sample to individuals aged 50 and above, excluding those with missing information on health related variables, thus, the final sample size is of 25,672 individuals.

The dependent variable is SRH, derived from the question ‘How is your health in general?’ with five possible answers: ‘very good’, ‘good’, ‘fair’, ‘bad’, ‘very bad’. We categorised this variable into three groups: 0 for ‘very good’ and ‘good’; 1 for ‘fair’; and 2 for ‘bad’ and ‘very bad’.

Independent variables include age groups (‘50–54’, ‘55–59’, ‘60–64’, ‘65–69’, ‘70–74’, ‘75 + ’), marital status (‘single’, ‘married’, and ‘divorced or widow’), education (‘no education or primary’, ‘lower secondary’, ‘upper secondary’, ‘tertiary’), employment status (‘retired’, ‘employed’, ‘unemployed’, ‘housekeeper’ ‘inactive’), migration background (‘native’, ‘foreign-born’), and macro-area of residence in Italy (‘North’, ‘Centre’, ‘South’) (Table 1).

3 Methods

In many disciplines data are frequently misclassified. In the statistical literature, there are several models that address misclassification in binary regression [5, 13–15], but there are few models that consider measurement error for polychotomous responses. Misclassification of a dependent variable Y means that an observation with a true value j is observed as k . This mistake could easily happen, for example, during an interview if the respondent misunderstands the question or the interviewer simply checks the wrong box.

Ignoring the presence of misclassification is not trivial; in fact, when traditional estimation methods (like ordered logit or probit) are used in discrete-choice contexts with a misclassified dependent variable, the resulting estimates are inconsistent.

Table 1. Sample percentage distribution for each individual characteristic by sex.

	Men	Women
Age groups		
50–54	17.82	16.20
55–59	16.49	15.03
60–64	14.89	13.88
65–69	13.91	13.19
70–74	12.99	12.26
75 +	23.90	29.44
Marital status		
Single	11.39	8.51
Married	76.02	58.77
Divorced/Widow	12.59	32.72
Education		
No education/Primary	22.82	33.16
Lower secondary	34.27	28.94
Upper secondary	30.80	27.72
Tertiary	12.12	10.17
Employment status		
Retired	52.39	35.24
Employed	39.58	24.27
Unemployed	4.67	3.07
Housekeeper	0.18	33.66
Inactive	3.19	3.76
Migration background		
Native	95.87	93.63
Foreign-born	4.13	6.37
Macro-area of residence		
North	44.78	44.55
Centre	19.44	19.36
South	35.77	36.09
Outcome		
<i>Self-rated health</i>		
Very good/Good	58.18	49.52

(continued)

Table 1. (continued)

	Men	Women
Fair	30.76	34.71
Poor/Very poor	11.06	15.77
<i>N. observations</i>	<i>11,810</i>	<i>13,862</i>

Source: Authors’ elaboration on IT-EHIS (wave 3) data.

To obtain consistent estimates of the regression coefficients β , we need to explicitly take into account the misclassification probabilities $\Pr(Y^O = j|Y^T = k) = \alpha_{jk}$.

In particular, the probability of observing j conditionally on a set of independent variables is:

$$\Pr(Y^O = j|X) = \sum_{k=1}^k a_{jk} \Pr(Y^T = k|X) \tag{1}$$

where K are the categories of the response variable. We then model $\Pr(Y^T = k|X)$ via a cumulative ordered logit model as:

$$\Pr(Y^T \leq k|X) = \Lambda(\gamma_k - X\beta) \tag{2}$$

where $\Lambda(\cdot)$ is the logistic cumulative density function.

The vector of parameters $\theta = (\beta, \gamma, \alpha)$ is estimated through likelihood maximisation.

For the purpose of our study, we conducted two separate analyses. We first fit a standard ordered logit model based on Eq. (2) only, and then an adjusted one base on both Eqs. (1) and (2), comparing the result.

All analyses are stratified by sex.

4 Expected Findings

In line with the literature on misclassification [16, 17], we expect the association between the outcome variable and the independent variables will be weakened in the naive model in case of misclassification.

Additionally, consistent with previous findings about SRH, we expect higher levels of education to correlate with better SRH as well as being in a stable relationship to have a protective effect.

This study addresses the subjectivity and potential misclassification in SRH. By applying adjusted models, we aim to obtain more accurate estimates in understanding SRH and its relationship with various predictors, contributing to better population health assessment.

Acknowledgments. We acknowledge co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE8 – Project Age-It: “Ageing Well in an Ageing Society”. This resource was co-financed by the Next Generation EU [DM 1557 11.10.2022]. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

1. Lundberg, O., Manderbacka, K.: Assessing reliability of a measure of self-rated health. *Scand. J. Soc. Med.* **24**(3), 218–224 (1996)
2. Idler, E.L., Benyamini, Y.: Self-rated health and mortality: a review of twenty-seven community studies. *J. Health Soc. Behav.* 21–37 (1997)
3. Jylhä, M.: What is self-rated health and why does it predict mortality? Towards a unified conceptual model. *Soc. Sci. Med.* **69**(3), 307–316 (2009)
4. Sen, A.: Health: perception versus observation: self-reported morbidity has severe limitations and can be extremely misleading. *BMJ* **324**(7342), 860–861 (2002)
5. Hausman, J., Abrevaya, J., Scott-Morton, F.: Misclassification of the dependent variable in a discrete-response setting. *J. Economet.* **87**(2), 239–269 (1998)
6. Zajacova, A., Dowd, J.B.: Reliability of self-rated health in US adults. *Am. J. Epidemiol.* **174**(8), 977–983 (2011)
7. Jürges, H.: True health vs response styles: exploring cross-country differences in self-reported health. *Health Econ.* **16**(2), 163–178 (2007)
8. Crossley, T.F., Kennedy, S.: The reliability of self-assessed health status. *J. Health Econ.* **21**(4), 643–658 (2002)
9. Salomon, J.A., Tandon, A., Murray, C.J.: Comparability of self-rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ* **328**(7434), 258 (2004)
10. Bowling, A., Windsor, J.: The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA). *J. Epidemiol. Commun. Health* **62**, 81–85 (2008)
11. Fakhoury, J., Burton-Jeangros, C., Guessous, I., Consoli, L., Duvoisin, A., Jackson, Y.: Self-rated health among undocumented and newly regularized migrants in Geneva: a cross-sectional study. *BMC Public Health* **21**, 1198 (2021)
12. Perneger, T.V., Gayet-Ageron, A., Courvoisier, D.S., Agoritsas, T., Cullati, S.: Self-rated health: analysis of distances and transitions between response options. *Qual. Life Res.* **22**, 2761–2768 (2013)
13. Naranjo, L., Martín, J., Perez, C.J., Rufo, M.J.: Addressing misclassification for binary data: probit and t-link regressions. *J. Stat. Comput. Simul.* **84**, 2187–2213 (2014)
14. Arezzo, M.F., Guagnano, G.: Misclassification in binary choice models with sample selection. *Econometrics* **7**(3), 32 (2019)
15. Arezzo, M.F., Arima, S., Guagnano, G.: A two-part measurement error model to estimate participation in undeclared work and related earnings. *Statist. Model.* (2023)
16. Neuhaus, J.M.: Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* **86**(4), 843–855 (1999)
17. Liu, H., Zhang, Z.: Logistic regression with misclassification in binary outcome variables: a method and software. *Behaviormetrika* **44**, 447–476 (2017)



Changes in Parenthood Expectancy Over Time Under an Age-Perspective

Alessandra Trimarchi¹✉ and Erich Striessnig²

¹ University of Messina, Messina, Italy
alessandra.trimarchi@unime.it

² University of Vienna, Vienna, Austria

Abstract. Fertility postponement and increases in life expectancy have contributed to changes in parenthood expectancy (PEX), i.e., the time people can expect to live as an ever-parent. While previous work showed gender and country differences in PEX, nothing is known about age-specific contributions to changes in PEX. We fill this gap by using data from the Human Fertility Database (HFD) and the Human Mortality Database (HMD), and applying decomposition analysis based on a stepwise replacement algorithm. We find that age-specific positive contributions to changes in PEX between 1970 and 2019 mostly derive from the life-expectancy component. Focusing on reproductive ages (15–49), we observe that losses in PEX due to the fertility component conceal some of those gains. This does not hold at very late reproductive ages, though. Understanding these age-patterns can inform us on the specific sub-populations contributing to changes in the indicator over time, which then can have applications in individual-level analyses.

Keywords: Fertility · Mortality · Parental Lifespan · Decomposition

1 Introduction

Unprecedented increases in the age at first birth, and increases in life expectancy have been global tendencies since the second half of the 20th century. These trends have had consequences for family relations, because of the lifetime shared by parents and their children. Family relations are relevant to understand societal dynamics, in particular with regard to intergenerational support [1, 2]. Parents invest in their offspring and support them well into adult life. In turn, children may become a support for their parents during old age. These reciprocal, supportive relations may not establish successfully between all parents and their offspring, however, in most cases the longer parents support their children, the better the outcomes for the children (e.g., in terms of educational attainment and socio-economic status) and, vice versa, the better the outcomes for parents later on in life (e.g., in terms of active ageing and better health status) [3, 4].

To monitor changes in intergenerational support over time across national contexts, “Parenthood Expectancy” (PEX) has recently been suggested as a period-specific indicator synthesizing fertility and mortality information [5]. PEX is defined as the time

people can expect to live as an ever-parent. In particular, it depends on the mean age at first birth (MAFB), and the remaining life expectancy at that age in a certain period. Across countries, most recent PEX estimates were found to vary between 50–58 years for women and tended to be up to 10 years shorter for men. The population of interest contributing to PEX are those individuals entering parenthood in a given year irrespective of age. This implies that each individual may contribute differently to the aggregate indicator: the younger the individual becomes a parent, the sooner in their lifetime they start contributing person-years to overall PEX, the later the entry into parenthood, the less remaining life years to contribute. Still, due to marked increases in latest-late fertility [6], and constant improvements in health at older ages [7], late reproductive ages may have become key in determining changes in PEX. This, obviously, varies by context.

In this paper, we add to previous work by presenting age-specific contributions to changes in PEX over time, allowing us to shed light on fertility and longevity dynamics in conjunction. In particular, rather than assuming all years of parenthood to be equivalent, e.g., with respect to parents' support capacities, we are interested in gains and losses of PEX at different ages. Previous work has shown that changes in fertility and mortality behaviors pretty much cancel each other out over time in determining PEX, especially for mothers [5]. However, this does not have to be the case when looking at age-specific contributions to PEX and gains at some ages might be compensated for by losses in other age brackets. Understanding these age-patterns can inform us on the specific populations contributing to changes in the indicator over time, which then can have applications in individual-level analyses, e.g. if subjectively created beliefs about PEX affect people's fertility behavior, a question that remains surprisingly understudied.

To analyze age-specific changes in PEX over time, we use harmonized fertility and mortality data of 12 countries. Since information about age-specific fertility rates is not available for men, here we focus only on women. Furthermore, as both fertility postponement and mortality improvements are rather slow-moving processes, we look at those countries with relatively long time series. Methodologically, we apply decomposition analysis, and counterfactual analysis to study differences in the age-specific contributions to PEX over time.

2 Background

Since the 1970s, demographic changes occurring across industrialized countries have had structural consequences for the family. A combination of low mortality and fertility produced a change in intergenerational relations: within each family, more generations are likely to coexist today than in the past. But, because of fertility decline, each generation tends to be smaller than the previous one [2].

Changes in the intensity and duration of the time that different generations spend together can play an important role also in shaping demographic behavior [8]. Parental support for the child through adulthood is considered a key factor in the child's fertility [9]. At older ages, parents benefit from investments in their children [10], which can improve health status, and consequently life expectancy [4].

Age at first birth and life expectancy at this age affect the duration of time spent as an ever-parent in a certain period [5]. People becoming a parent at young age have a higher life expectancy than older parents, and are therefore contributing the most person-years to PEX. However, with the increasing prevalence of late first births, i.e., births at age 40 and over [6], more women can contribute to PEX at late reproductive ages. We expect that, while gains in PEX derive mostly from improvements in life expectancy long after reproductive ages, gains in PEX occurring during reproductive ages are concealed by fertility postponement at late reproductive ages.

3 Data and Methods

Our analysis builds on period data from the Human Fertility Database (HFD) for women across a range of developed countries. 12 of these provide times series of sufficient length to make comparisons over the 50 years starting in 1970. Additionally, we use life table information from the Human Mortality Database (HMD). From these period life tables, we extract life expectancies at the mean ages at first birth indicated by HFD. Since HFD reports MAFB in decimals, whereas remaining life years are available only at complete ages, we round MAFB to the closest integer and subtract/add the difference from/to the remaining life years at the closest complete age. Thus, implicitly we assume that life expectancy declines linearly between single years of age.

Our decomposition analysis of PEX is based on the *stepwise replacement* algorithm described in Andreev et al. [11] as implemented in the *DemoDecomp* package of the R Statistical Software [12]. For every age, it allows us to decompose the overall change in PEX between two time points into its MAFB- and e(x)-components. Formally:

$$PEX^{(2019)} - PEX^{(1970)} = \sum_{x=15}^{\omega} \delta_x \quad (1)$$

where δ_x represents the difference in PEX by age x between the two periods.

$$\delta_x = \frac{1}{2} \times (\delta_x^{2019-1970} - \delta_x^{1970-2019}) \quad (2)$$

4 Results and Conclusions

Figure 1 shows the age-specific decomposition of PEX by its components, fertility and life expectancy. PEX has generally increased in all countries since 1970, and these positive gains are a consequence of improvements in life expectancy at older ages, i.e. children born in 2019 can expect to spend more life years with their elderly parents than those born in 1970. During reproductive ages, though, improvements in longevity do not compensate for losses deriving from fertility postponement, reflecting that more first-born children are born to parents that are already in late reproductive ages. Still, at very late reproductive ages (roughly from around age 40), across all countries improvements in longevity can compensate for losses in PEX derived from very late entry into parenthood.

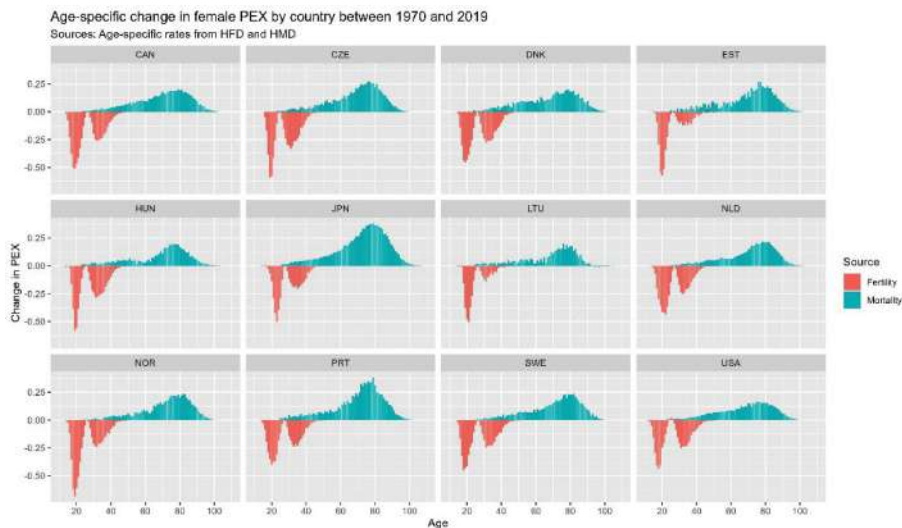


Fig. 1. Age-specific change in female PEX by country between 1970 and 2019.

Only in Lithuania, we observe gains in PEX after age 40 coming from the fertility component. This is slightly counterintuitive at first, as MAFB has also increased in Lithuania in parallel with total fertility rates declining. Further inspection of Lithuanian data focusing on age-specific fertility rates (ASFRs) by parity between 1970 and 2019, we find that after age 30 women in Lithuania were having their first births (ASFR1) at a higher rate in 2019 than in 1970 due to first-birth postponement. This implies that fertility was more *polarized* in 1970, because fewer women were having way more kids than in 2019, when fertility declined, but more women were uniformly having at least one kid. Given that only ASFR1 matters for PEX, more women could contribute parenthood years in late reproductive ages in 2019 than back in 1970.

Changes in PEX are the result of several underlying and interacting processes. For instance, improvements in health, including reproductive health, and progresses in assisted reproductive technology, which contributed to the increasing prevalence of late motherhood in more recent periods.

Figure 1 focuses on the comparison between 1970 and the most recent available time point, and we might miss important changes in PEX that may have happened in the earlier part of this period. Therefore, for the few countries where sufficiently long time series are available, Fig. 2 shows the comparison between 1950 and 1990. This comparison reflects the strong East-West divide that characterized this time period. While the Western nations for which data is available (Canada, Netherlands and the USA) all show a pattern relatively similar to the one found in the 1970–2019 comparison, the available countries that were part of the Eastern Bloc (Bulgaria, Czech Republic and Slovak Republic) reveal a different pattern, with gains in PEX deriving from both mortality improvements and changes in the mean ages at first birth. This is due to marked differences in the onset of fertility postponement. While we are starting to see this happening in the Western countries already in the 1970ies, Eastern countries

experienced postponement only after the fall of the Iron Curtain. As a consequence, in these countries women in 1990 contributed more person-years to PEX than women in 1950 almost at all reproductive ages.

In a next step, we apply counterfactual analyses, and compare the empirical, age-specific baseline to a hypothetical country-specific scenario, which is created by combining the lowest MAFB ever observed for a country with the highest life expectancy observed in that country, possibly in a different year.

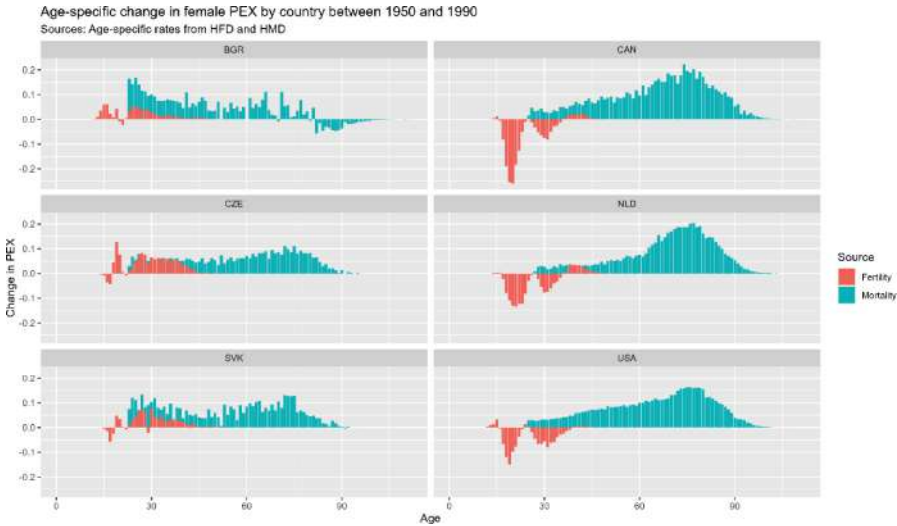


Fig. 2. Age-specific change in female PEX by country between 1950 and 1990.

References

1. Kalmijn, M.: Parents' socioeconomic status and support to adult children across the life course. *J. Marriage Fam.* **86**, 7–29 (2024). <https://doi.org/10.1111/jomf.12928>
2. Hagestad, G.O.: Transfers between grandparents and grandchildren: The importance of taking a three-generation perspective. *J. Family Res.* **18**, 315–332 (2006). <https://doi.org/10.20377/jfr-298>
3. Jiang, N.: Adult children's education and later-life health of parents in China: the intergenerational effects of human capital investment. *Soc. Indic. Res.* **145**, 257–278 (2019). <https://doi.org/10.1007/s11205-019-02109-9>
4. Sabater, A., Graham, E., Marshall, A.: Does having highly educated adult children reduce mortality risks for parents with low educational attainment in Europe? *Ageing Soc.* **1–36** (2019). <https://doi.org/10.1017/s0144686x19000795>
5. Striessnig, E., Trimarchi, A.: How much time is left? International trends in parenthood expectancy. *Demogr. Res.* **48**, 421–438 (2023)
6. Beaujouan, E.: Latest-late fertility? Decline and resurgence of late parenthood across the low-fertility countries. *Popul. Dev. Rev.* **46**, 219–247 (2020). <https://doi.org/10.1111/padr.12334>

7. Chernew, M., Cutler, D.M., Ghosh, K., Landrum, M.B.: Understanding the improvement in disability free life expectancy in the US elderly population. *Natl. Bureau Econ. Res.* (2016)
8. Bengtson, V.L.: Beyond the nuclear family: the increasing importance of multigenerational bonds. *J. Marriage Fam.* **63**, 1–16 (2001). <https://doi.org/10.1111/j.1741-3737.2001.00001.x>
9. Rutigliano, R.: Counting on potential grandparents? Adult children's entry into parenthood across european countries. *Demography* **57**, 1393–1414 (2020). <https://doi.org/10.1007/s13524-020-00890-8>
10. Fingerman, K.L., Birditt, K.S.: Chapter 14 - relationships between adults and their aging parents. In: Schaie, K.W., Willis, S.L. (eds.) *Handbook of the Psychology of Aging*, 7th edn., pp. 219–232. Academic Press, San Diego (2011)
11. Andreev, E.M., Shkolnikov, V., Begun, A.: Algorithm for decomposition of differences between aggregate demographic measures and its application to life expectancies, healthy life expectancies, parity-progression ratios and total fertility rates. *Demogr. Res.* **7**, 499–522 (2002)
12. Riffe, T.: Package 'DemoDecomp' decompose demographic functions (2018)



Bayesian Latent Factor Model for Multi-target Inference

Bruno Ursino¹(✉), Manuela Antonelli², Beatrice Cantoni², Ilenia Epifani¹,
and Francesco Trovò³

¹ Dipartimento di Matematica, Politecnico di Milano, Milan, Italy
`bruno.ursino@mail.polimi.it`, `ilenia.epifani@polimi.it`

² Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milan, Italy
`{manuela.antonelli,beatrice.cantoni}@polimi.it`

³ Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano,
Milan, Italy
`francesco1.trovo@polimi.it`

Abstract. Providing safe drinking water is essential for human health. To achieve this, analyzing multidimensional spectroscopic measurements of water, such as absorbance and fluorescence spectra, is crucial. However, measuring such parameters entails significant time and instrumental efforts. Therefore, it is essential to develop statistical tools to minimize the required lab analyses. We propose a bivariate functional data model, where each sampling unit comprises the bivariate target: absorbance and fluorescence. We employed a Bayesian bivariate functional latent factor model extending [4]. In our water analyses application, interpretable posterior distributions of the latent factors are crucial, thus we addressed their identifiability question, using Varimax-RSP method. Hamiltonian Monte Carlo sampler was applied to sample the posterior distribution. We developed a Python package, available on GitHub, implementing our model and we tested it on simulated univariate and bivariate data, anticipating its application with real data from the European Project SafeCREW [1].

Keywords: Bayesian latent factor model · Bivariate functional data · Identifiability of latent factors · Water NOM characterization

1 Introduction and Motivating Example

Natural organic matter (NOM) plays a critical role in aquatic ecosystems and directly impacts water quality, influencing aspects from taste and smell to the risk of harmful disinfection byproducts (DBPs) like trihalomethanes and haloacetic acids. In fact, DBPs are generated when NOM interacts with chemical disinfectants used in water treatment, causing potential health hazards. Moreover, it was proved that different NOM compositions generate different proportions of DBPs and, thus, risks. Consequently, accurately quantifying and

understanding NOM's concentration and composition are crucial steps to predict DBPs generation, effectively control water treatment processes and ensure the safety of drinking water. The characterization of NOM through UV-Visible spectroscopy involves analysis of the multivariate functional data collected for each water sample [3]. Usually, water NOM is measured by analyses of: absorbance spectra, which are relatively easier to acquire, and the more complex, time consuming and expensive fluorescence spectra. Data generated by both absorbance and fluorescence spectroscopy can be represented as functions across a two-dimensional domain, adding layers of complexity to the spectroscopic analysis. On the other hand, given the intricate nature of these spectroscopic measurements, among the different models developed to offer clear interpretability of fluorescence, the Parallel Factor Analysis (PARAFAC) technique stands out in its ability to deconstruct fluorescence spectra into a series of factors. These factors are conceptualized as tensor products of curves, thereby streamlining the interpretation of complex multivariate functional data. This research is motivated by the need to quantify the relationship between absorbance and fluorescence spectra, across various scenarios characterized by different values of covariates. Examples of covariates often accessible include information on the treatment applied to the sample, such as the contact time in case of treatment through adsorption on Activated Carbon. Understanding this relationship is crucial for devising effective methods to predict NOM characteristics using more readily accessible measurements.

We propose a bivariate Bayesian latent factor (BBLF) model built on the Bayesian regression model for a univariate functional target outlined in [4]. Our extension allows this model to encompass a multivariate framework capable of analyzing multiple spectroscopic measurements. While our methodology shares similarities with the bivariate extension presented in [2], which focuses on imputing missing data, our primary objective is to address the identifiability of a set of latent factors shared by two functional targets. These factors are intricately linked to each sample, and the rationale is to ascertain common latent variables that describe the impact of different solutes on the spectra. This is a simulation study anticipating application with real drinking water data.

2 Bivariate Bayesian Latent Factor Model Specification

The proposed Bayesian model extends previous work on univariate functional data models by incorporating a multivariate approach that allows for the simultaneous analysis of multiple spectroscopic measurements.

Likelihood. The likelihood of the BBLF model is depicted schematically in Fig. 1. Each of the two targets functional data vector $\mathbf{Y}^{(\iota)}$, measured in L_ι locations for $\iota = 1, 2$, is modeled as a Gaussian random vector centered around a linear combination of a set of target ι -specific basis functions $\mathbf{B}^{(\iota)} = (B_1^{(\iota)}, \dots, B_{p_\iota}^{(\iota)})$, with corresponding target-specific basis coefficients $\boldsymbol{\theta}^{(\iota)} = (\theta_1^{(\iota)}, \dots, \theta_{p_\iota}^{(\iota)})^T$, $\iota = 1, 2$. The $\boldsymbol{\theta}^{(\iota)}$ are linearly regressed onto a small set of latent factor scores $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^T$ with ι -specific loading factors $\boldsymbol{\Lambda}_j^{(\iota)} = (\Lambda_{j1}^{(\iota)}, \dots, \Lambda_{jk}^{(\iota)})^T$, $j = 1, \dots, p_\iota$; the $\boldsymbol{\eta}$ common to $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ is centered on a linear combination of r covariates $\boldsymbol{\beta}\mathbf{X}$. The large number of bases p_1, p_2 makes the BBLF model flexible, allowing for a wide range of potential functional shapes, while the small length k of $\boldsymbol{\eta}$ helps to reduce the model dimension.

Given a bivariate random sample $(\mathbf{Y}_1^{(1)}, \mathbf{Y}_1^{(2)}), \dots, (\mathbf{Y}_N^{(1)}, \mathbf{Y}_N^{(2)})$ and the covariate values recorded for the s th unit $\mathbf{X}_s = (X_{1s}, \dots, X_{rs})^T$, the contribution of the s th pair $(\mathbf{Y}_s^{(1)}, \mathbf{Y}_s^{(2)})$ to the likelihood is given by:

$$\begin{aligned} \mathbf{Y}_s^{(\iota)} | \boldsymbol{\theta}_s^{(\iota)}, \psi^{(\iota)} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{B}^{(\iota)} \boldsymbol{\theta}_s^{(\iota)}, \psi^{(\iota)} \mathbf{I}), \\ \boldsymbol{\theta}_{j,s}^{(\iota)} | \boldsymbol{\Lambda}_j^{(\iota)}, \boldsymbol{\eta}_s, \tau_{\theta_j}^{(\iota)} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\eta}_s^T \boldsymbol{\Lambda}_j^{(\iota)}, 1/\tau_{\theta_j}^{(\iota)}), \quad \boldsymbol{\eta}_s | \boldsymbol{\beta} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\beta} \mathbf{X}_s, \mathbf{I}). \end{aligned} \quad (1)$$

In addition, marginalizing Eq. (1) w.r.t. $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ yields the following representation of Eq. (1):

$$\begin{aligned} \mathbf{Y}_s | \Theta &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N} \left(\begin{bmatrix} \mathbf{B}^{(1)} \boldsymbol{\Lambda}^{(1)} \boldsymbol{\beta} \mathbf{X}_s \\ \mathbf{B}^{(2)} \boldsymbol{\Lambda}^{(2)} \boldsymbol{\beta} \mathbf{X}_s \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \quad s = 1, \dots, N, \\ \Sigma_{\iota\iota} &= \psi^{(\iota)} \mathbf{I} + \mathbf{B}^{(\iota)} \left[\Delta_{\theta}^{(\iota)} + \boldsymbol{\Lambda}^{(\iota)} \boldsymbol{\Lambda}^{(\iota)T} \right] \mathbf{B}^{(\iota)T}, \\ \Sigma_{12} &= \mathbf{B}^{(1)} \boldsymbol{\Lambda}^{(1)} \boldsymbol{\Lambda}^{(2)T} \mathbf{B}^{(2)T}, \quad \Sigma_{21} = \mathbf{B}^{(2)} \boldsymbol{\Lambda}^{(2)} \boldsymbol{\Lambda}^{(1)T} \mathbf{B}^{(1)T} \end{aligned} \quad (2)$$

where $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{\tau}_{\theta}^{(1)}, \boldsymbol{\tau}_{\theta}^{(2)}, \psi^{(1)}, \psi^{(2)}\}$. Equation (2) highlights the additive structure of the heteroscedasticity, indicating that it is managed through the basis function coefficients. Furthermore, Eq. (2) allows for directly estimating the conditional distribution of one target given the other, which is one of the goals of this research.

Prior. To finalize the Bayesian model, we carefully selected suitable marginal priors for each parameter in the likelihood, starting from those appearing in the

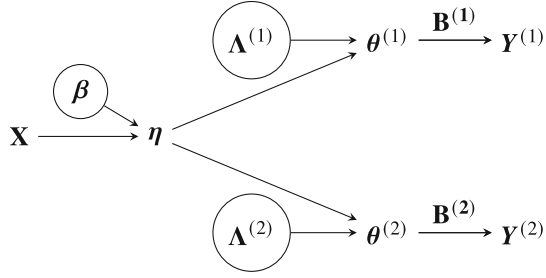


Fig. 1. Diagram of the BBLF model.

last block to those in the first one in Eq. (1), and assumed independence between blocks of parameters.

As the $k \times r$ regression coefficients β link the r covariates to the k latent factors, a priori they follow an uninformative *standard* student's t_χ distribution, allowing some coefficients to diverge from the mean, if supported by the data significantly.

For the loadings $\Lambda^{(1)}, \Lambda^{(2)}$ capturing the latent structure of the functional targets, we assume that a priori they follow i.i.d. multiplicative gamma process shrinkage priors as $\Lambda^{(\iota)} \sim \text{MGPS}(\nu, \alpha_1, \alpha_2)$, i.e., for $j = 1, \dots, p_\iota$ and $\ell = 1, \dots, k$:

$$\Lambda_{j,\ell}^{(\iota)} | \tau_{\lambda,\ell}^{(\iota)} \stackrel{\text{i.i.d.}}{\sim} t_\nu \left(0, \tau_{\lambda,\ell}^{(\iota)} \right), \quad \tau_{\lambda,\ell}^{(\iota)} = \prod_{\gamma=1}^{\ell} \delta_\gamma^{(\iota)}, \quad \delta_1^{(\iota)} \sim \text{Ga}(\alpha_1, 1), \quad \delta_\ell^{(\iota)} \sim \text{Ga}(\alpha_2, 1) \quad \ell \geq 2.$$

The parameter $\tau_{\lambda,\ell}^{(\iota)}$ is the precision of the ℓ th loading, and works as a shrinkage parameter for η_ℓ . The hyperparameters α_1, α_2 modulate the regularization degree: if $\alpha_2 > 2$ the prior predictive's total variance grows to a finite limit as the number of latent factors increases, as one can see from the variance formula $\mathbb{V}\text{ar} \left[\Lambda_{j,\ell}^{(i)} \right] = \nu / (\nu - 2) \times (\alpha_2 - 1)^{1-\ell} / (\alpha_1 - 1)$. Adopting this prior enables effective basis selection [4].

The number of latent factors k is unknown, and an optimal value can be determined by comparing models with varying k using posterior predictive checks. For the p_ι precisions $\tau_{\theta,j}^{(\iota)}$ of each basis j and target ι we assume a priori independence and the usual Gamma distribution: $\tau_{\theta,j}^{(\iota)} \sim \text{Ga}(\alpha_\sigma, \beta_\sigma)$; each of them accommodates heteroscedasticity in the functional representation of the targets, providing a non-vanishing likelihood for residual functions that the latent factors failed to capture. Finally, the variances $\psi^{(1)}, \psi^{(2)}$ of the measurement noises of $\mathbf{Y}_1, \mathbf{Y}_2$ are i.i.d. $\text{Inv-Ga}(\alpha_\psi, \beta_\psi)$ -distributed.

3 Posterior Inference

Posterior Regression Analysis of the Two Functional Targets: Methodology. The relationship between new observed targets $\tilde{\mathbf{Y}}^{(1)}$ and $\tilde{\mathbf{Y}}^{(2)}$, across various scenarios identified by new covariates $\tilde{\mathbf{X}}$, is addressed by regressing one target, for instance $\tilde{\mathbf{Y}}^{(1)}$, on the other, $\tilde{\mathbf{Y}}^{(2)}$; in a Bayesian framework, this is provided by the posterior predictive distribution of $\tilde{\mathbf{Y}}^{(2)}$ given $\tilde{\mathbf{Y}}^{(1)}$. To obtain this, we make use of the conditional distribution of $\tilde{\mathbf{Y}}^{(2)}$ given $\tilde{\mathbf{Y}}^{(1)}$ and the parameter set Θ , and from Equation (2) we derive that:

$$\tilde{\mathbf{Y}}^{(2)} | \Theta, \tilde{\mathbf{Y}}^{(1)} \sim \mathcal{N} \left(\mathbf{B}^{(2)} \Lambda^{(2)} \beta \tilde{\mathbf{X}} + \Sigma_{21} \Sigma_{11}^{-1} [\tilde{\mathbf{Y}}^{(1)} - \mathbf{B}^{(1)} \Lambda^{(1)} \beta \tilde{\mathbf{X}}], \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right). \quad (3)$$

Finally, we marginalize Eq. (3) w.r.t. the posterior distribution of Θ , given the data. In practice, the posterior distribution of Θ is approximated using MCMC, and therefore, the posterior predictive function is obtained by averaging over the posterior samples of Θ . Furthermore, analyzing the regression from a broader perspective, the posterior distribution of the composite parameter $\Sigma_{21} \Sigma_{11}^{-1}$ allows inference on the portion of $\mathbf{Y}^{(1)}$ pertinent for predicting $\mathbf{Y}^{(2)}$.

Identifiability of the Latent Factors: Methodology. A fundamental challenge in interpreting the latent factors $\boldsymbol{\eta}$ within our BBLF model arises from their non-identifiability, as the $\boldsymbol{\eta}$ are equivalent under any rotation and signed permutation. To tackle this issue, we employ the Varimax-RSP (Rotational Sign Permutation) algorithm, introduced in [5]. This algorithm identifies an optimal rotation based on the Varimax criterion and ensures consistency across posterior samples, by applying an appropriate signed permutation to each. The Varimax-RSP algorithm is applied to the matrix of loading factors $\boldsymbol{\Lambda}$, which is formed by concatenating the two target-specific matrices $\boldsymbol{\Lambda}^T = [\boldsymbol{\Lambda}^{(1)T} \ \boldsymbol{\Lambda}^{(2)T}]$. The methodology enables the projection of latent factors post-Varimax-RSP adjustment, considering that, conditional on the model parameters Θ , the latent variables $\boldsymbol{\eta}$ are multivariate normal distributed, with the following mean and covariance matrices:

$$\mathbb{E} [\tilde{\boldsymbol{\eta}} | \Theta, \tilde{\mathbf{Y}}] = \beta \tilde{\mathbf{X}} + \begin{bmatrix} \mathbf{B}^{(1)} \boldsymbol{\Lambda}^{(1)} \\ \mathbf{B}^{(2)} \boldsymbol{\Lambda}^{(2)} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\mathbf{Y}}^{(1)} - \mathbf{B}^{(1)} \boldsymbol{\Lambda}^{(1)} \beta \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}}^{(2)} - \mathbf{B}^{(2)} \boldsymbol{\Lambda}^{(2)} \beta \tilde{\mathbf{X}} \end{bmatrix},$$

$$\text{Cov} [\tilde{\boldsymbol{\eta}} | \Theta, \tilde{\mathbf{Y}}] = \begin{bmatrix} \mathbf{B}^{(1)} \boldsymbol{\Lambda}^{(1)} \\ \mathbf{B}^{(2)} \boldsymbol{\Lambda}^{(2)} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}^{(1)} \boldsymbol{\Lambda}^{(1)} \\ \mathbf{B}^{(2)} \boldsymbol{\Lambda}^{(2)} \end{bmatrix}.$$

These equations enable us to draw samples from the posterior predictive distribution of $\tilde{\boldsymbol{\eta}}$ for any new observation $\tilde{\mathbf{Y}}$, similarly to the process outlined for the posterior predictive distribution given by $\tilde{\mathbf{Y}}_i^{(2)} | \tilde{\mathbf{Y}}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}$. This approach maps the new observations onto the latent space and provides a measure of the uncertainty associated with these projections via the posterior distribution.

Posterior Computation. Monte Carlo sampling from the posterior distribution was conducted using the Hamiltonian Monte Carlo (HMC) method via the Stan [6] software, with initial chain states determined through Laplace sampling. A Python package implementing such a procedure has been made publicly accessible on GitHub [7].

The computational efficiency of the likelihood calculation within each HMC iteration is significantly enhanced by projecting the data $\mathbf{Y}^{(\iota)}$ onto the space of basis functions, particularly because $L_\iota \gg p_\iota$. If the basis matrix $\mathbf{B}^{(\iota)}$ is of full column rank, we use the decomposition:

$$\|\mathbf{Y}^{(\iota)} - \mathbf{B}^{(\iota)} \boldsymbol{\theta}^{(\iota)}\|_2^2 = \|\mathbf{R}^{(\iota)} (\mathbf{B}^{(\iota)+} \mathbf{Y}^{(\iota)} - \boldsymbol{\theta}^{(\iota)})\|_2^2 + \|\mathbf{M}^{(\iota)} \mathbf{Y}^{(\iota)}\|_2^2$$

where $\mathbf{B}^{(\iota)+}$ is the Moore-Penrose pseudo-inverse, $\mathbf{M}^{(\iota)} = \mathbf{I} - \mathbf{B}^{(\iota)} \mathbf{B}^{(\iota)+}$ and $\mathbf{R}^{(\iota)}$ matrix comes from the thin-QR decomposition of $\mathbf{B}^{(\iota)}$.

Simulation Study. We simulated a dataset by sampling from the prior distribution of the BBLF model. We employed 10 cubic b-splines for $\mathbf{Y}^{(1)}$ and truncated Fourier basis (third harmonic) with a linear trend for $\mathbf{Y}^{(2)}$, setting $k = 4$ latent factors. Covariates (5 per sample) followed a normal distribution

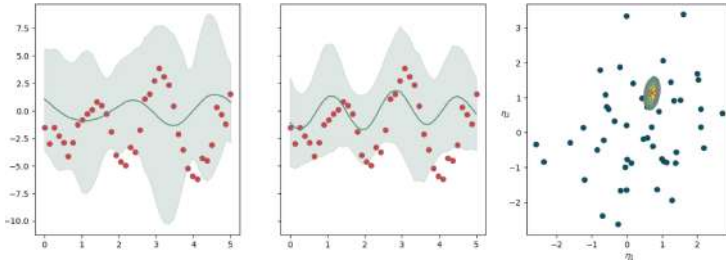


Fig. 2. The solid lines of left and center plots show the posterior predictive distributions $\hat{\mathbf{Y}}^{(2)} | \tilde{\mathbf{X}}, \mathbf{Y}$ and $\hat{\mathbf{Y}}^{(2)} | \tilde{\mathbf{X}}, \hat{\mathbf{Y}}^{(1)}, \mathbf{Y}$, respectively; the dots are the sampled values for $\hat{\mathbf{Y}}^{(2)}$. (Right) Projection of the first two latent factors of 50 samples of $\hat{\mathbf{Y}}^{(2)}$; the shaded area shows the credibility region of $\hat{\boldsymbol{\eta}} | \tilde{\mathbf{X}}, \hat{\mathbf{Y}}^{(1)}, \hat{\mathbf{Y}}^{(2)}, \mathbf{Y}$ for the sample plotted on the first two plots.

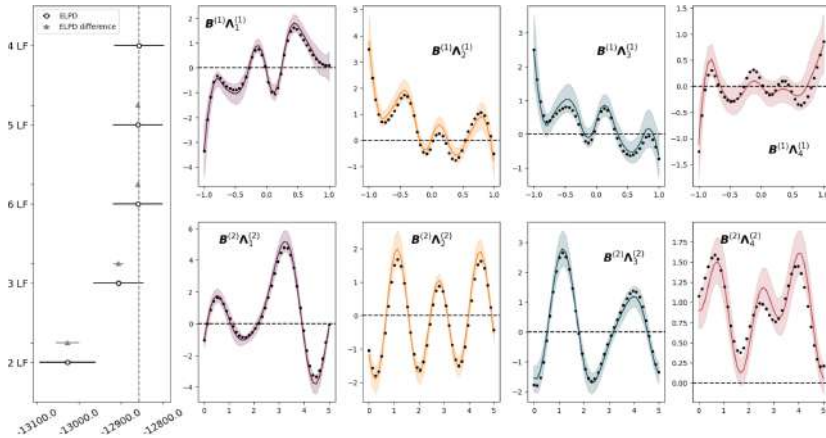


Fig. 3. (Left) ELPD for BBLF models with $k = 2, \dots, 6$ and their differences w.r.t. $k = 4$. (Right) Solid lines show posterior means, and dots are true values of the loading factors post-Varimax; top graphs represent $\mathbf{B}^{(1)}\boldsymbol{\Lambda}^{(1)}$; bottom graphs show $\mathbf{B}^{(2)}\boldsymbol{\Lambda}^{(2)}$.

($\sigma = 0.3$). Observations for $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ were made on 40 evenly spaced points over the intervals $(-1, 1)$ and $(0, 5)$, respectively. We set $\alpha_1 = 2.1$, $\alpha_2 = 2.5$, $\nu = \chi = 4$. We decided to focus on datasets with one-dimensional domains for both target variables, to simplify the presentation of our model and enhance clarity in interpreting the output. However, it is important to point out that the BBLF model is capable of handling more complex scenarios, including those with target variables spanning multi-dimensional spaces. These capabilities are particularly relevant to the nuanced demands of spectroscopic analysis in NOM characterization. For readers interested in exploring deeper insights or more complex examples, supplementary material is available on the GitHub repository [7].

In terms of regression analysis, Fig. 2 shows an example of posterior predictive distribution of $\mathbf{Y}^{(2)}$ on a test set of 50 samples, using both covariates alone and also incorporating the other target variable $\mathbf{Y}^{(1)}$. In this case, we estimate the RMSE:

$$\text{RMSE}[\tilde{\mathbf{Y}}^{(2)}] = \sigma[\tilde{\mathbf{Y}}^{(2)}] \simeq 3.9, \text{RMSE}[\tilde{\mathbf{Y}}^{(2)}|\tilde{\mathbf{X}}] \simeq 2.93, \text{RMSE}[\tilde{\mathbf{Y}}^{(2)}|\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}^{(1)}] \simeq 2.38.$$

Figure 2 and the reported RMSE values confirm that predictions improve as additional information is incorporated. Regarding the inference on the loading factors, the Pareto smoothed importance sampling leave-one-out method allows for comparisons among models with varying k . Notably, the estimated Expected Log pointwise Predictive Density (ELPD), shown on the leftmost subplot of Fig. 3, confirms $k = 4$ latent factors as optimal, consistently with the true model configuration. This validation is further supported by examining the varimaxed latent factors of models with $k > 4$, indicating no additional significant factors, see [7]. On the rightmost part of Fig. 3 we compare the estimated loading factors $\mathbf{B}^{(1)}\mathbf{\Lambda}^{(1)} | \mathbf{Y}$ with the true varimaxed counterparts (top), and $\mathbf{B}^{(2)}\mathbf{\Lambda}^{(2)} | \mathbf{Y}$ against their true varimaxed equivalents (bottom). It can be observed that most of the true values lie within the respective 90% posterior credibility intervals, corroborating the idea that the Varimax-RSP method can identify the true latent factors.

Acknowledgements. This work was supported by SafeCREW - Grant no. 101081980, founded by European Union - Horizon Europe, and MUR - Prin 2022 - Grant no. 2022CLTYP4, funded by the European Union - Next Generation EU.

References

1. Cantoni, B., et al.: SafeCREW: a holistic project to reduce NOM and DBPS in drinking water and minimize the human health risk under climate change scenarios. In: Gordon Research Conference (2023)
2. Jang, J., Manatunga, A., Chang, C., Long, Q.: Bayesian multiple imputation approach to bivariate functional data with missing components. *Stat. Med.* **40**, 4772–4793 (2021)
3. Matilainen, A., Gjessing, E., Lahtinen, T., Hed, L., Bhatnagar, A., Sillanpää, M.: An overview of the methods used in the characterisation of natural organic matter (NOM) in relation to drinking water treatment. *Chemosphere* **83**, 1431–1442 (2011)
4. Montagna, S., Tokdar, S., Neelon, B., Dunson, D.: Bayesian latent factor regression for functional and longitudinal data. *Biometrics* **68**, 1064–1073 (2012)
5. Papastamoulis, P., Ntzoufras, I.: On the identifiability of Bayesian factor analytic models. *Stat. Comput.* **32**(2), 23 (2022)
6. Stan Development Team: Stan Modeling Language Users Guide and Reference Manual (2024). <https://mc-stan.org>. 2.34
7. Ursino, B., Epifani, I., Trovò: Data analysis toolkit (2024). https://github.com/Addicted2BayesianEpistemology/bayesian_multitarget_latent_factors. GitHub Repository



Clustering Multivariate Rating Data Within the CUB Framework

Matteo Ventura^{1(✉)}, Julien Jacques², and Paola Zuccolotto¹

¹ Department of Economics and Management, University of Brescia, Brescia, Italy
`{matteo.ventura,paola.zuccolotto}@unibs.it`

² ERIC Laboratory, University Lyon 2, Lyon, France
`julien.jacques@univ-lyon2.fr`

Abstract. Among the models for the analysis of rating data, the CUB (Combination of discrete Uniform and shifted Binomial random variable) is particularly interesting because it gives an interpretation of the dual latent factors believed to influence the final decision of a rater: feeling and uncertainty. In essence, this model represents the distribution of final ratings as a combination of a shifted binomial and a uniform random variable. Within the framework provided by the CUB model, we propose a mixture of multivariate CUB models to cluster multivariate rating data, whose estimation is performed via the EM algorithm. To evaluate our approach, we conducted two simulations, showcasing the model's consistency in handling complex data structures and capturing underlying patterns. Our findings underscore the potential of this methodology in uncovering hidden structures within multivariate rating datasets, offering valuable insights in various research domains.

Keywords: Ordinal data · Model-based clustering · EM algorithm · Mixture model

1 The CUB Model

In marketing, psychological and social research, questionnaires serve as a common tool to evaluate latent traits such as the perceptions, opinions, and attitudes of respondents. These traits are often assessed using Likert-type rating scales, where the results are gathered as ordinal data, meaning that there is a natural ordering of the categories. Analyzing ordinal data can be challenging because it requires addressing the inherent properties of ordinal variables. To tackle this challenge, the CUB (Combination of discrete Uniform and shifted Binomial random variables) model has been introduced [2], and then extended by several researchers [6]. This model posits that the underlying decision-making process leading to respondents' final ratings is characterized by two latent components: the feeling and the uncertainty.

The feeling component represents the rational aspect, reflecting respondents' preferences for a particular item or attribute. On the other hand, the uncertainty component captures inherent indecision present in human choices.

Given a scale with m categories, the rating $r = 1, \dots, m$ is considered as the realisation of a mixture of a shifted Binomial and a discrete Uniform random variables, defined as follows:

$$P(r \mid \xi, \pi) = \pi P_{SB}(r \mid \xi) + (1 - \pi) P_U(m),$$

where $P_{SB}(r \mid \xi) = \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r}$, and $P_U(m) = \frac{1}{m}$. The former accounts for the feeling, measured by the so-called feeling parameter $1 - \xi \in [0, 1]$, and the latter accounts for the uncertainty, which is measured by the uncertainty parameter $1 - \pi \in (0, 1]$, that is, the mixing proportion of the mixture.

2 A Mixture of Multivariate CUB Models

Clustering is an important tool for researchers to discover hidden structures in data sets, and mixture models have been successfully applied to cluster data. However, ordinal data have received less attention compared to other types of data, especially in a model-based clustering context, where there are a few proposals. The two most notable models are a model-based clustering relying on a Binary Ordinal Search (BOS) algorithm [1] and clustMD, which assumes that ordinal data are a generalization of a latent Gaussian [5].

In this section, a mixture model for clustering multivariate rating data following the CUB paradigm is presented. The CUB model is univariate, however, clustering is usually performed with multivariate data. Therefore, we propose a clustering algorithm based on a mixture of multivariate CUB models, which are assumed to be conditionally independent.

Let $\mathbf{r} = (r_{ij})_{1 \leq i \leq n, 1 \leq j \leq J}$ be a multivariate ordinal variable, where the j th component is an ordinal variable with m_j categories; and let ω_k be the mixing proportion of cluster K , such that $\omega_k > 0$ and $\sum_{k=1}^K \omega_k = 1$.

Under the assumption of conditional independence of the variables, the marginal distribution of $\mathbf{r}$ is:

$$P(\mathbf{r} \mid \boldsymbol{\theta}) = \sum_{k=1}^K \omega_k \prod_{j=1}^J \left[\pi_{jk} P_{SB}(r_j \mid \xi_{jk}) + (1 - \pi_{jk}) P_U(m_j) \right],$$

with $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega})$, where $\boldsymbol{\xi} = (\xi_{jk})_{1 \leq j \leq J, 1 \leq k \leq K}$, $\boldsymbol{\pi} = (\pi_{jk})_{1 \leq j \leq J, 1 \leq k \leq K}$, $\boldsymbol{\omega} = (\omega_k)_{1 \leq k \leq K}$.

For mixture models there exists no close form for the ML estimators, therefore the parameters are estimated via the EM algorithm. To define the complete log-likelihood, two latent allocation variables are introduced in the model: $\mathbf{z} : (\mathbf{z}_k)_{1 \leq k \leq K}$ is a random variable distributed as a one-order Multinomial distribution, $\mathbf{z} \sim \mathcal{M}(1; \omega_1, \dots, \omega_K)$, such that $z_{ik} = 1$ if the i th rater preferences come from cluster k , and $z_{ik} = 0$ otherwise. The second allocation variable $\mathbf{v} : (\mathbf{v}_j)_{1 \leq j \leq J}$, is dependent on the allocation variable z_{ik} and it is distributed as a Bernoulli with parameter π_{jk} . The variable $v_{ij} = 1$ if the preference of the i th rater for the j th item comes from a Shifted Binomial of parameter ξ_{jk} , and

$v_{ij} = 0$ if it belongs to a Uniform random variable. Therefore, the complete log-likelihood is defined as follows:

$$\begin{aligned} \ell_c(\boldsymbol{\theta}; \mathbf{r}, \mathbf{z}, \mathbf{v}) = & \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left\{ \ln(\omega_k) + \sum_{j=1}^J v_{ij} [\ln(\pi_{jk}) + \ln [P_{SB}(r_{ij} | \xi_{jk})]] \right. \\ & \left. + \sum_{j=1}^J (1 - v_{ij}) [\ln(1 - \pi_{jk}) + \ln P_U(m_j)] \right\}. \end{aligned}$$

Starting from the initial set of parameters $\boldsymbol{\theta}^0 = (\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega})^{(0)}$, the t -th iteration ($t \geq 0$) is the following:

- **E-step:** for all $1 \leq i \leq n$, $1 \leq j \leq J$, $1 \leq k \leq K$, the conditional probabilities of $\mathbf{z}$ are computed as follows:

$$\begin{aligned} & \mathbb{E}(z_{ik} | \mathbf{r}_i; \boldsymbol{\theta}^{(t)}) \\ = & \frac{\omega_k^{(t)} \prod_j [\pi_{jk}^{(t)} P_B(r_{ij} | \xi_{jk}^{(t)}) + (1 - \pi_{jk}^{(t)}) P_U(m_j)]}{\sum_{k'=1}^K \omega_{k'}^{(t)} \prod_j [\pi_{jk'}^{(t)} P_B(r_{ij} | \xi_{jk'}^{(t)}) + (1 - \pi_{jk'}^{(t)}) P_U(m_j)]} \\ = & \tau_{ik}(\mathbf{r}_i; \boldsymbol{\theta}^{(t)}) = \tau_{ik}^{(t)}. \end{aligned}$$

Then, the conditional probabilities of the product $\mathbf{z} \cdot \mathbf{v}$ are computed:

$$\begin{aligned} & \mathbb{E}(z_{ik} v_{ij} | \mathbf{r}_i; \boldsymbol{\theta}^{(t)}) \\ = & \frac{\pi_{jk}^{(t)} P_B(r_{ij} | \xi_{jk}^{(t)})}{\pi_{jk}^{(t)} P_B(r_{ij} | \xi_{jk}^{(t)}) + (1 - \pi_{jk}^{(t)}) P_U(m_j)} \cdot \tau_{ik}(\mathbf{r}_i; \boldsymbol{\theta}^{(t)}) \\ = & \nu_{ik}(r_{ij}; \boldsymbol{\theta}^{(t)}) \cdot \tau_{ik}(\mathbf{r}_i; \boldsymbol{\theta}^{(t)}) = \eta_{ijk}^{(t)}. \end{aligned}$$

- **M-step:** update the estimation of the model parameters:

$$\begin{aligned} \pi_{jk}^{(t+1)} &= \frac{\sum_{i=1}^n \eta_{ijk}^{(t)}}{\sum_{i=1}^n \tau_{ik}^{(t)}}, \\ \xi_{jk}^{(t+1)} &= \frac{\sum_{i=1}^n \eta_{ijk}^{(t)} (m_j - r_{ij})}{\sum_{i=1}^n \eta_{ijk}^{(t)} (m_j - 1)}, \\ \omega_k^{(t+1)} &= \frac{\sum_{i=1}^n \tau_{ik}^{(t)}}{n}. \end{aligned}$$

The algorithm is stopped when a threshold is reached in the relative change of the log-likelihood.

3 Simulation Studies

In this section, a simulation study is performed to evaluate the performance of the EM algorithm. 100 datasets of size $n \in \{100, 1000\}$ have been simulated from a trivariate mixture ($J = 3$) of two CUB components ($K = 2$), each dimension having nine categories ($m_j = 9$). The set of parameters considered to generate the simulated data sets is shown in Table 1.

Table 1. Set of parameters chosen for generating the simulated data sets.

	$k = 1$			$k = 2$		
ω	0.40			0.60		
	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
π	0.60	0.60	0.50	0.70	0.80	0.70
ξ	0.30	0.20	0.10	0.70	0.80	0.70

3.1 Simulation Study 1

On each data set, the EM algorithm is run 20 times with uniform random starting parameters and setting the number of clusters equal to the real number ($K = 2$). The algorithm stops when the relative change in the log-likelihood is less than 10^{-4} . Finally, the run which provided the higher log-likelihood among the 20 has been considered as the final output of the EM algorithm. The results improve as the number of observations increases. The estimation of the parameter π_{jk} seems to be more difficult compared to the estimation of the other parameters. The boxplots of the parameter estimates distributions with a sample of size $n = 100$ and $n = 1000$ are reported in Figs. 1, 2 and 3.

3.2 Simulation Study 2

The second simulation aims to check if the model can detect the real number of clusters. As in the previous simulation, the EM algorithm is run 20 times on both the data sets with uniform random initialization, and it is stopped when the threshold on the relative change in the log-likelihood is reached. The EM algorithm is run by considering different numbers of clusters $K \in \{1, \dots, 5\}$, with $K = 1$ meaning that there are no clusters and the model is simply a multivariate CUB. The estimated number of clusters corresponds to the one that leads to the lower BIC value. The performances in the detection of the real number of clusters are good with both sample sizes since the model detects two clusters in 97% of cases with the data set of size $n = 100$, and in 100% of cases in the data set of size $n = 1000$.

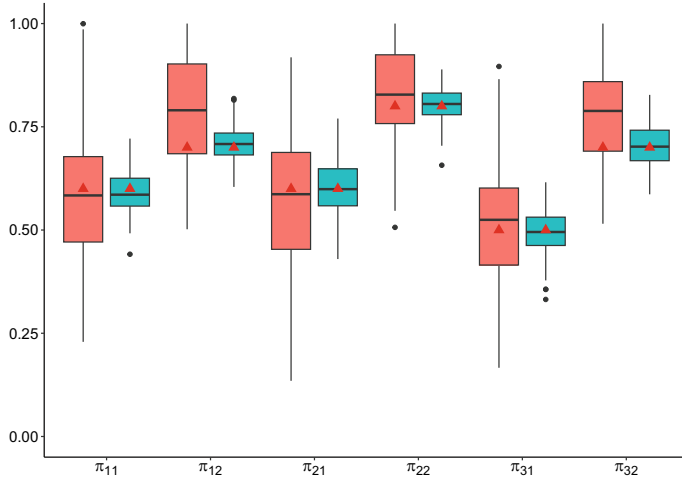


Fig. 1. Boxplots of the estimates of the parameters π_{jk} considering the sample with $n = 100$ observations (in red), and $n = 1000$ observations (in blue). The red triangles represent the true values of the parameters.

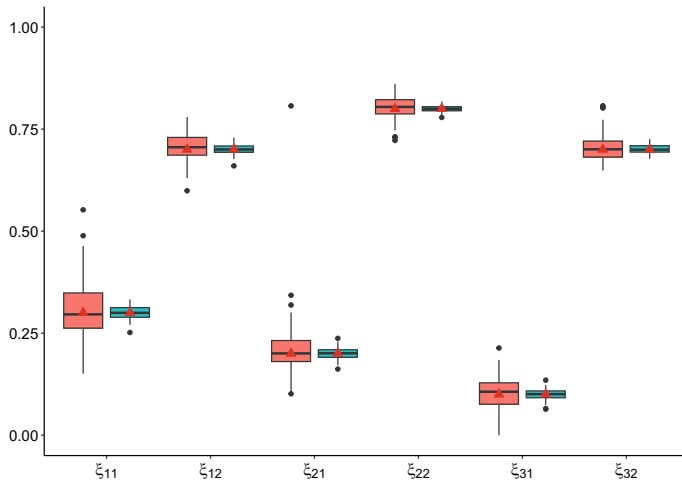


Fig. 2. Boxplots of the estimates of the parameters ξ_{jk} considering the sample with $n = 100$ observations (in red), and $n = 1000$ observations (in blue). The red triangles represent the true values of the parameters.

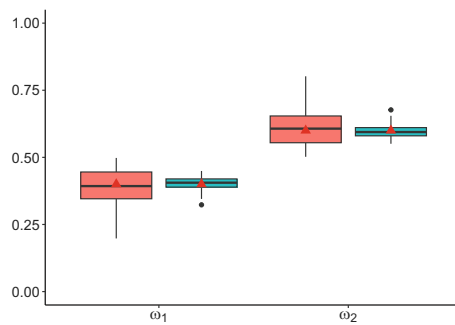


Fig. 3. Boxplots of the estimates of the parameters ω_k considering the sample with $n = 100$ observations (in red), and $n = 1000$ observations (in blue). The red triangles represent the true values of the parameters.

4 Conclusions

In this study, we introduced a mixture model for clustering rating data within the CUB framework. The simulation results demonstrate promising outcomes in estimating the mixing proportion and the feeling parameter. However, estimating the uncertainty parameter appears to be more challenging. Additional simulation scenarios will be explored to further assess the model's performance.

It is important to note that a mixture of CUB models is not identifiable due to the presence of the uniform component of different CUB models, which are not distinguishable [1, 4]. However, we expect that the probability of a non-identifiable solution decreases as the number of variables in the data increases. Moreover, we intend to apply the model to real datasets to demonstrate its practical utility.

References

1. Biernacki, C., Jacques, J.: Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Stat. Comput.* **26**, 929–943 (2016)
2. D'Elia, A., Piccolo, D.: A mixture model for preferences data analysis. *Comput. Stat. Data Anal.* **49**(3), 917–934 (2005)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* **39**(1), 1–38 (1977)
4. Grilli, L., Iannario, M., Piccolo, D., Rampichini, C.: Latent Class CUB models. *Adv. Data Anal. Classif.* **8**(1), 105–119 (2014)
5. McParland, D., Gormley, I.C.: Model based clustering for mixed data: clustMD. *Adv. Data Anal. Classif.* **10**(2), 155–169 (2016). <https://doi.org/10.1007/s11634-016-0238-x>
6. Agresti, A., Kateri, M.: The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Stat. Methods Appl.* **28**(3), 445–449 (2019). <https://doi.org/10.1007/s10260-019-00468-8>



Fairness-Adaptive Classification Trees

Anna Gottard¹, Sabrina Giordano², and Vanessa Verrina²(✉)

¹ Department of Statistics, Computer Science, Applications “G. Parenti”, University of Florence, Firenze, Italy

anna.gottard@unifi.it

² Department of Economics, Statistics and Finance “Giovanni Anania”, University of Calabria, Arcavacata, Italy

{sabrina.giordano,vanessa.verrina}@unical.it

Abstract. Artificial intelligence and machine learning algorithms are increasingly driving many decisions that can impact several aspects of human life, including healthcare, college admissions, recruitment, and loan provision, among others. Given their extensive influence on our daily experiences, it is imperative to adopt algorithms that are aware of the uncertainty in their predictions and of the discrimination that they can induce. This article explores various methods for quantifying uncertainty in fairness metrics. Furthermore, it introduces a modification to the CART algorithm designed to improve fairness in binary classification tasks by incorporating uncertainty in these metrics. This modified algorithm adaptively applies penalties to unfair splits, thereby enhancing prediction fairness while preserving a high level of accuracy.

Keywords: Algorithm bias · CART algorithm · Fairness-aware machine learning · Uncertainty

1 Introduction

Over the past decade, Machine Learning (ML) algorithms have become widely used in decision-making contexts, notably in areas with significant societal impacts. Consequently, there is a growing concern regarding their potential to perpetuate discrimination against specific groups of people based on sensitive attributes like gender, race and religion. Various factors contribute to the unfairness observed in ML algorithms. Firstly, biases entrenched within training datasets present a considerable challenge. These biases often originate from diverse sources, such as inaccurate device measurements, historically biased human decisions, or erroneous data entries. Importantly, ML algorithms tend to replicate and magnify these biases, leading to the perpetuation of unfair outcomes. Secondly, biases may arise in the sample selection mechanism. Such an issue results in datasets that inadequately represent the target population, possibly further exacerbating unfairness. Additionally, biases can stem from the inherent objectives of ML algorithms, which prioritize the reduction of the overall prediction errors. This objective-centric bias may inadvertently favour majority

groups over minorities, thereby reinforcing systemic inequities. Moreover, biases might emerge from the reliance on *proxy* attributes instead of directly using sensitive attributes in decision-making processes.

While sensitive attributes like race, gender, and age are typically considered inappropriate for decision-making, proxy attributes (non-sensitive attributes indirectly linked to sensitive attributes) can inadvertently perpetuate bias when included in datasets. These distinct categories underscore the complex nature of unfairness within ML systems and emphasize the necessity of comprehensive bias mitigation strategies. Therefore, alongside enhancing the performance of a machine learning algorithm, the emerging task for data scientists is to assess whether the model predictions exhibit discrimination and to devise strategies for mitigating such undesirable bias to the greatest extent possible. Great attention must be given to the trade-off between accuracy and fairness because prioritizing a higher degree of fairness may compromise accuracy [1].

Classification and Regression Trees (CART) [2], and related algorithms, provide a transparent, white-box approach to prediction and classification. The inherent transparency of CART allows practitioners to derive predictions directly from the tree visual representation without the need to use a computer. Unfortunately, a notable drawback is the algorithm’s susceptibility to discrimination bias, a consequence of its greedy nature that can lead to an over-reliance on the background variables [3].

In this paper, we propose a variation of the CART algorithm that pushes towards fairer predictions. There are three main approaches in ML to promote fairness. The first approach, *pre-processing*, involves the elimination or modification of biased elements within the training data to prevent discriminatory influences. The second strategy, *post-processing* entails adjusting the model predictions post-training to ensure alignment with established fairness metrics, often through recalibration of outputs or altering decision thresholds. Lastly, the *in-processing* technique directly integrates fairness considerations into the algorithm learning objectives during its training phase. This integration can be realized by introducing regularization terms or constraints tailored to promote fairness, or by embedding a penalty term into the objective function to penalize unfair predictions, as detailed in the work of [4]. Our work is situated within the realm of *in-processing* techniques for discrimination attenuation. In particular, we propose to penalize the loss function in classification trees, taking into account the uncertainty in the fairness metric. The splits providing a confidence interval of the fairness metric not including zero are penalized by a factor that adapts to the amount of unfairness. Notice that, we will limit to the case of the response variable and the sensible attribute are binary. In this case, we can use standard confidence intervals for comparing proportions. The only assumption needed for utilizing the proposed algorithm is the independence and identical distribution of the sampled units.

The paper is organized as follows. In Sect. 2 we present some of the most used fairness metrics together with their confidence intervals. Section 3 describes the

fairness-adapted CART algorithm and its performances on data. The paper ends with some concluding remarks (Sect. 4).

2 Fairness Metrics and Their Uncertainty Evaluation

Fairness in automated decision-making systems is commonly understood as the absence of any discriminatory bias in their outcomes. However, the concept of fairness is inherently multifaceted and context-dependent. For a comprehensive overview of fairness metrics, see, among many, [5].

In this section, we present the prevalent fairness metrics along with an approach to evaluate their uncertainty. We consider a classical binary supervised classification problem with a random sample $\{\mathbf{X}_i, S_i, Y_i\}_{i=1}^n$, where $\mathbf{X}_i$ is the feature vector with p predictors of the i th unit, S_i is the sensitive attribute and Y_i is the binary response variable. In particular, we assume that $S_i = 1$ for the privileged group and $S_i = 0$ for the unprivileged group. Moreover, $\hat{Y}_i$ is the predicted value for Y_i with 1 denoting the positive outcome.

2.1 Some Metrics for Fairness

According to the following metrics, a classification is considered fair if the proportion of the positive predictions is similar across groups of the sensitive attribute.

- **Statistical Parity (SP)**. This metric, also known as *Demographic Parity*, compares the probabilities of positive predictions across groups by a difference, as

$$\Delta_{SP} = |P[\hat{Y} = 1 \mid S = 1] - P[\hat{Y} = 1 \mid S = 0]|. \quad (1)$$

A lower value of this measure indicates more similar acceptance rates and therefore better fairness [4].

- **Disparate Impact (DI)**. This metric compares the probabilities of positive predictions across groups by a ratio, as

$$\Delta_{DI} = \frac{P[\hat{Y} = 1 \mid S = 1]}{P[\hat{Y} = 1 \mid S = 0]}, \quad (2)$$

where a value of the metric close to 1 denotes the absence of discrimination [6].

- **Equalized Odds (EO)**. This metric compares the statistical parity separately for positive and negative outcomes, therefore distinguishing the false-positive rates (FPRs) and the true-positive rates (TPRs)

$$\begin{aligned} \Delta_{EO-FPR} &= |P[\hat{Y} = 1 \mid S = 1, Y = 0] - P[\hat{Y} = 1 \mid S = 0, Y = 0]|, \\ \Delta_{EO-TPR} &= |P[\hat{Y} = 1 \mid S = 1, Y = 1] - P[\hat{Y} = 1 \mid S = 0, Y = 1]|, \end{aligned} \quad (3)$$

where a value of the metrics close to 0 indicates the absence of discrimination [4].

2.2 Uncertainty Evaluation

The measures of fairness reported above are ratios or differences in probabilities that are typically estimated from sample data and thus affected by uncertainty, which is often overlooked. An estimate of $P[\hat{Y} = 1 \mid S = 1]$ is the sample proportion of positive predictions in the privileged group, and similarly, an estimate of $P[\hat{Y} = 1 \mid S = 0]$ is the corresponding sample proportion in the unprivileged group. Confidence Intervals (CIs) for these quantities allow sampling uncertainty to be taken into account.

Let n_{11} and n_{10} be the observed frequencies of the positive predictions in the privileged group and the unprivileged group whose sample sizes are n_1 and n_0 , respectively. Let $\hat{p}_{11} = n_{11}/n_1$ and $\hat{p}_{10} = n_{10}/n_0$ denote the corresponding estimated proportions. The traditional asymptotic Wald $(1 - \alpha)\%$ confidence interval for the difference of two proportions is

$$\hat{p}_{11} - \hat{p}_{10} \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_{11}(1 - \hat{p}_{11})}{n_1} + \frac{\hat{p}_{10}(1 - \hat{p}_{10})}{n_0}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. This CI can be calculated for the Statistical Parity and for the Equalized Odds. Regarding the latter, the preceding symbols must be referred not only to the privileged group but also to the true value of Y , so for example n_{111} would denote the observed frequencies of the positive predictions ($\hat{Y} = 1$) in the privileged group ($S = 1$) with true positive value ($Y = 1$).

One of the simplest interval for the ratio of proportions is the *log* interval, adjusted to avoid issues related to zero frequencies

$$\log \frac{(n_{11} + 0.5)/(n_1 + 0.5)}{(n_{10} + 0.5)/(n_0 + 0.5)} \mp z_{\alpha/2} \sqrt{\frac{1}{n_{11} + 0.5} + \frac{1}{n_{10} + 0.5} - \frac{1}{n_1 + 0.5} - \frac{1}{n_0 + 0.5}}.$$

These intervals are just the simplest and fastest to compute. More sophisticated and better-performing intervals have been proposed in the literature. An overview of proposed confidence intervals for difference and ratios of proportions for both small and large samples is in [7] and a recent review on score-test based CIs is proposed by [8].

3 A Fairness-Adapted CART Algorithm

In this section, we will describe our proposed *in-processing* technique for discrimination attenuation. The main idea is to introduce a penalty to the Information Gain (IG) within the traditional CART algorithm that is linked to the CI of a fairness metric measured on each possible split. This penalty takes into account both the amount of unfairness and the uncertainty associated with it. In particular, we use the standard confidence intervals proposed in Sect. 2.2 to compare

proportions. Splits that yield CIs for the fairness metric that do not include zero are subjected to a penalty factor in IG as follows

$$IG_{\text{FAIR}} = \begin{cases} IG & \text{if } 0 \in \text{CI}\{\Delta\} \\ \lambda(1 - \phi_{\text{CI}\{\Delta\}})IG & \text{if } 0 \notin \text{CI}\{\Delta\} \end{cases} \quad (4)$$

where $\lambda \in [0, 1]$ is a tuning parameter controlling the trade-off between accuracy and fairness, ϕ is the minimum distance of the confidence interval from zero, Δ is the chosen fairness metric and CI its confidence interval. The proposed approach is schematized in Algorithm 1.

Algorithm 1. Fairness-Adapted CART Algorithm

Require: Data D , Sensitive Attribute S , Fairness metric Δ , Penalty factor λ

- 1: Initialize an empty decision tree T
 - 2: Build the decision tree T using CART algorithm where
 - 3: **for** each potential split in T **do**
 - 4: Compute the fairness metric Δ for the split its CI
 - 5: **if** CI of Δ does not include zero **then**
 - 6: Calculate the minimum distance of the CI from zero, $\phi_{\text{CI}\{\Delta\}}$
 - 7: Apply penalty to the IG by multiplying $\lambda(1 - \phi_{\text{CI}\{\Delta\}})$ to IG
 - 8: **end if**
 - 9: **end for**
 - 10: **Output:** Decision tree T with *fairness* adapted information gain
-

As an initial demonstration, we applied our Fair-Adapted CART algorithm on synthetic data created from the diamond DAG, as described in [3], adapted for a binary classification task. This data generation involves a nonlinear relationship between features and the target variable. The sensitive attribute is the background variable, set as binary. For fairness evaluation, we considered the Statistical Parity metric, Δ_{SP} (Eq. (1)). We further partitioned the training data into train and validation sets using random splitting. The validation subset was used to obtain honest estimates of accuracy and fairness for several values of λ . The optimal value of λ for the synthetic data was determined to be 0.1. Accuracy of the algorithm under this selected setting was assessed as the proportion of correct predictions on the validation set. These performance results are shown in Table 1. Our approach significantly enhances the fairness of the model, reaching a CI that crosses 0. However, as expected, this comes at the expense of a decreased accuracy.

Table 1. Validation set performance and fairness measures comparison for $\lambda = 0.1$.

	Population	CART	Fair CART
Δ_{SP}	0.17	0.26	0.11
CI	$(-0.02, 0.36)$	$(0.08, 0.45)$	$(-0.02, 0.25)$
Accuracy	—	0.58	0.52

4 Concluding Remarks

In conclusion, the pervasive integration of artificial intelligence and machine learning algorithms across various domains raises critical concerns regarding fairness. The impact of these algorithms extends to crucial aspects of human life, such as healthcare, education, employment, and finance. Acknowledging the potential for discrimination inherent in these systems, it becomes imperative to develop and deploy algorithms that not only optimize accuracy but also mitigate bias. In this context, we have proposed a modification of the CART algorithm, the Fairness-Adapted CART algorithm, aimed at enhancing fairness in classification tasks while considering fairness uncertainty. We introduced a mechanism to penalize splits that yield fairness metric confidence intervals excluding zero. By incorporating uncertainty into the fairness assessment, Fairness-Adapted CART aims to strike a balance between accuracy and fairness while considering uncertainty, crucial in real-world applications where trade-offs between these metrics are inevitable.

References

1. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, pp. 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2017)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. CRC Press (1984)
3. Gottard, A., Vannucci, G., Marchetti, G.M.: A note on the interpretation of tree-based regression models. *Biom. J.* **62**(6), 1564–1573 (2020)
4. Pessach, D., Shmueli, E.: A review on fairness in machine learning. *ACM Comput. Surv.* **55**(3), 1–44 (2022)
5. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**(5), 1–42 (2018)
6. Castelnovo, A., Cosentini, A., Malandri, L., Mercorio, F., Mezzanzanica, M.: FFTree: a flexible tree to handle multiple fairness criteria. *Inf. Process. Manag.* **59**(6) (2022)
7. Fagerland, M., Lydersen, S., Laake, P.: Statistical Analysis of Contingency Tables. CRC Press (2017)
8. Agresti, A., Giordano, S., Gottard, A.: A review of score-test-based inference for categorical data. *J. Quant. Econ.* **20**(Suppl 1), 31–48 (2022)



An Ensemble Method for Disease Surveillance

Anna Vesely^(✉), Giulia Roli, Michele Scagliarini, and Rossella Miglio

Department of Statistical Sciences, University of Bologna, Bologna, Italy

anna.vesely2@unibo.it

Abstract. Healthcare heavily relies on effective disease surveillance, particularly during epidemiological emergencies. Predicting trends is challenging but crucial for effective policies. We propose an ensemble modelling framework to forecast time series of counts, such as the number of cases or hospitalizations. The method relies on multiple models, including flexible regressions with cubic splines, generalized additive models, Bayesian structural time series, and integer-valued generalized autoregressive conditional heteroscedasticity models. Predictions are computed as weighted averages on some of the individual models' predictions, where model selection and weights are based on performance on past observations. This approach allows to exploit diverse model strengths and compensate for individual weaknesses, ensuring more stable predictions. The behavior of the proposed method is illustrated on daily COVID-19 hospitalization counts in Emilia-Romagna in 2022, with a focus on periods showing changing trends.

Keywords: disease surveillance · ensemble · forecasting · COVID-19

1 Introduction

Disease surveillance is a crucial task in healthcare and life sciences, especially during periods of a health emergency, such as those occurred in the recent pandemic. A good surveillance system supports decisions towards an optimal allocation of human, technical and economic resources, allows the development of better and innovative health policies, and prevents critical situations during waves of emergency. Under this framework, forecasting trends is a challenging objective, as it can anticipate a potential change and allows to support better decision-making processes towards early and more effective health policy solutions.

In this work, we consider the problem of predicting time series of counts, such as the number of cases or hospitalizations. Several works on infectious diseases and other fields showed that ensemble approaches, which exploit and combine predictions from multiple models, are often more accurate than any forecast from a single model (see, e.g., [2]). Ensemble methods combine the strengths of different models, while compensating for the weaknesses of individual models

and producing more robust predictions. Indeed, if one model performs poorly in a certain scenario, its impact can be mitigated by the others in the ensemble.

We propose a general framework for an ensemble modelling approach, which provides predictions computed as a weighted average of different models' predictions. Candidate models are first analyzed in terms of performance on past observations, i.e., on how precisely they predict the last available observations. The accuracy of this prediction determines if and with which weight the model contributes to the ensemble. The method allows for any suitable choice of the prediction window and to include eventual covariates.

The behavior of the proposed approach is examined in the study of daily counts of COVID-19 hospitalizations in Emilia-Romagna. This analysis is the result of a long experience conducted on these data in collaboration with the Local Health Agency of Bologna, which suggested the importance of daily predictions based on a weekly frequency due to the organisational, managerial and executive set-up of hospitals. The effectiveness of an ensemble approach in this setting was illustrated in [6].

The manuscript is organized as follows. We give the general structure of the proposed ensemble approach in Sect. 2, then we introduce the contributing models in Sect. 3. The application on the COVID-19 data is shown in Sect. 4.

2 Ensemble Model

Suppose that we are interested in studying a count time series $y_t \in \mathbb{N}$, with $t = 1, \dots, T$, possibly in presence of a vector of covariates $\mathbf{x}_t \in \mathbb{R}^n$. The goal of the analysis is forecasting the next Q counts, i.e., computing predictions $\hat{y}_t$ for $t = T + 1, \dots, T + Q$. We assume that the predictions are calculated using data from the last W available time points ($t = T - W + 1, \dots, T$), where W is chosen according to the nature of the problem.

Each count y_t may be regarded as the realization of a Poisson random variable Y_t with parameter μ_t , where μ_t can be modelled in different ways. We propose an ensemble approach that takes into account multiple models and determines the desired predictions $\hat{y}_t$ as weighted averages of the individual models' predictions. The main idea is to fit a number K' of reasonable models, select $K \leq K'$ of them based on their performance on past observations, and finally compute

$$\hat{y}_t = \sum_{k=1}^K w_k \hat{y}_t^{(k)} \quad (t = T + 1, \dots, T + Q) \quad (1)$$

where $\hat{y}_t^{(k)}$ is the prediction given by the k -th model, and $w_1, \dots, w_K \in [0, 1]$ are some weights such that $\sum w_k = 1$.

The K selected models and the weights w_k are determined by shifting the focus back by Q time points. This way, each individual model is used to predict the last Q available counts ($t = T - Q + 1, \dots, T$), and then these predictions are compared to the observed data. Higher weights are given to the models that make the most accurate predictions, while those with the worst performance are

discarded and do not contribute to the ensemble. The pipeline of the proposed analysis is as follows.

1. **Fitting.** Fit all K' candidate models on the last W data points ($t = T - W + 1, \dots, T$). For any model choice, if different settings (e.g., pre-specified parameters) are possible, select the settings that lead to the lowest Bayesian Information Criterion (BIC) when appropriate.
2. **Selection.** Consider the last Q available counts. Fit all K' models (with the eventual settings as determined in the previous point) on $t = T - W - Q + 1, \dots, T - Q$. For each model, compute the predictions $\hat{y}_t^{(k)}$, $t = T - Q + 1, \dots, T$, and compare them to the observed counts y_t through the Mean Squared Deviation (MSD). Keep the K models that lead to the lowest MSD.
3. **Weight calibration.** Consider the last Q available counts and the selected models. Determine the weights that minimize the sum of squared differences between the ensemble predictions $\hat{y}_t$ and y_t , $t = T - Q + 1, \dots, T$. The minimization may be carried out through the Augmented Lagrangian method for equality constraints, implemented in the R package `nloptr`.
4. **Prediction.** Apply the selection from 2. and the weights from 3. to the models in 1. to calculate the desired predictions $\hat{y}_t$, $t = T, \dots, T + Q$.

Confidence intervals can be computed analogously to the ensemble predictions in (1), taking a weighted average of the bounds given by the selected models.

3 Contributing Models

In our proposal, we take into account a set of $K' = 7$ candidate models, $K = 5$ of which are then selected for the ensemble. The first, which we will refer to as Flexible [1], is a Generalized Linear Model (GLM) where the logarithm of the Poisson parameter μ_t is taken as a function of the covariates $\mathbf{x}_t$ and restricted cubic splines of the time t . The GLM and the splines function can be implemented using the R packages `stats` and `rms`, respectively.

The second model that is considered is the Generalized Additive Model (GAM) [5], which is similar to the Flexible but allows for a more general smooth function of time. It is implemented in the `mgcv` package.

Subsequently, we take the Bayesian Structural Time Series (BSTS) model [7], available in the package `bsts`. This method combines Kalman filtering to structure the time series, spike-and-slab regression to select regression components, and Bayesian model average to convert samples from repeated Markov Chain Monte Carlo (MCMC) draws into the final outputs. The model takes into account the covariates $\mathbf{x}_t$ and an autoregressive component.

The final block of models is identified in the INteger-valued Generalized AutoRegressive Conditional Heteroscedasticity (INGARCH) models [3, 4], where the parameter μ_t depends on the covariates $\mathbf{x}_t$, past observations $y_{t-1}, \dots, y_{t-p}$, past expectations $\mu_{t-1}, \dots, \mu_{t-q}$, and powers of time $t^0, t^1, \dots, t^r$. In particular, we choose four models, having $r = 0$ (Stationary), $r = 1$ (Linear), $r = 2$

(Quadratic), and $r = 3$ (Cubic). This family of models is implemented in the `tscount` package. If no prior information is available, the parameters corresponding to autoregression p and moving average q can be chosen as suggested in point 1. of the pipeline, considering all possible combinations of $p, q \in \{1, \dots, Q\}$ and identifying the lowest BIC.

4 Application to COVID-19 Hospitalizations

We examine the performance of the proposed ensemble method on data from the Local Health Agency on the daily number of hospitalizations for COVID-19, available at <https://github.com/pcm-dpc/COVID-19>. In particular, we focus on the non-intensive care unit (non-ICU) hospitalizations in Emilia-Romagna in 2022. The goal of the analysis is forecasting the counts on one week ($Q = 7$), using data from the previous three weeks ($W = 21$). As covariate, we take a factor variable that describes the day of the week (Sunday, Monday, etc.); holidays are coded as Sundays.

In disease surveillance interest usually lies in quickly detecting a change of trend, therefore we investigate if the method is able to catch such changes by comparing its predictions with the observed counts in four given periods. We examine two weeks where the number of hospitalizations starts to increase (Period 1: March 22–28, Period 3: June 21–27) or decreases after a peak (Period 2: May 3–9, Period 4: August 2–8). Figure 1 shows the time series under analysis, highlighting the considered time periods.

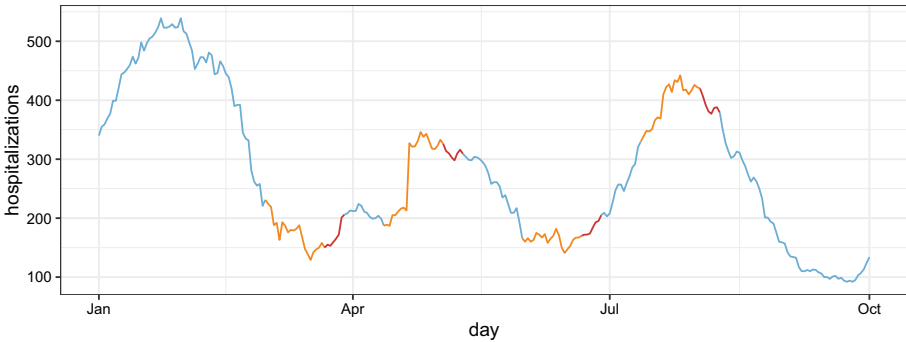


Fig. 1. Number of daily non-ICU hospitalizations for COVID-19 in Emilia-Romagna in 2022. The analysis focuses on four time periods: for each, the data from three weeks (orange) are employed to forecast the number of hospitalizations in the next week (red). (Color figure online)

Predictions given by the ensemble method, as well as the $K = 5$ models contributing to it, are shown in Fig. 2. Furthermore, Table 1 displays the MSD used in the model selection step (i.e., shifting the focus back by one week) and

the MSD computed on the prediction week of interest. We observe that the performance of individual models and the selection for the ensemble vary significantly over the four considered periods. However, the GAM is always selected and tends to provide accurate predictions. The ensemble predictions are, by construction, an average of the others; this mitigates extreme behaviours (see, e.g., the INGARCH cubic in Period 3) and tends to produce more stable results. The performance is particularly promising in Periods 2 and 3, while the method is not able to capture the newly increasing trend in Period 1.

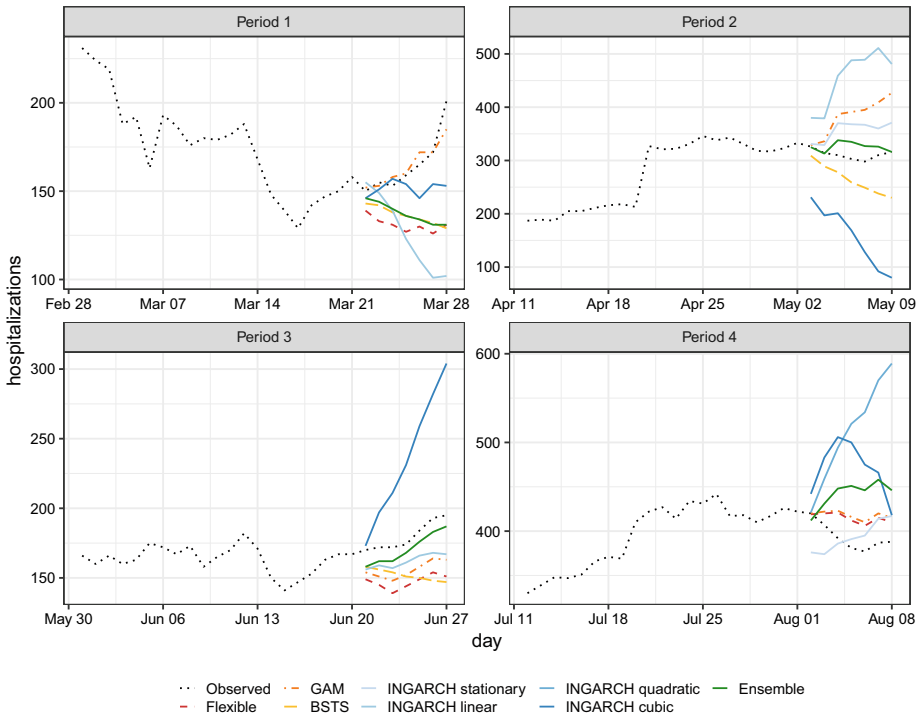


Fig. 2. Number of hospitalizations in four time periods: observed time series over four weeks, and predictions from different models in the last week. For each period, only the proposed ensemble model and the five individual models that contribute to the ensemble are displayed.

Future analyses may consider additional covariates other than the day of the week, such as the temperature and levels of air pollutants. Moreover, further research is needed on the comparison between the individual models and the ensemble. A high variability, e.g., expressed in terms of Coefficient of Variation (CV), may identify a critical situation, where the single models tend to be discordant, and thus the predictions become unstable, uncertain and less reliable. Preliminary analyses show that the highest CV values are obtained when the

Table 1. MSD between observed data and predictions of different models, in four time periods, for the selection step and the prediction. For each period, the two models with the highest selection MSD (in gray) are discarded.

Model	Period 1		Period 2		Period 3		Period 4	
	sel.	pred.	sel.	pred.	sel.	pred.	sel.	pred.
Flexible	23.2	23.2	59.2	10.4	18.7	18.6	14.8	6.2
GAM	22.0	4.2	21.7	26.0	9.9	13.8	17.6	7.0
BSTS	11.6	21.4	17.4	16.7	9.2	17.2	23.1	13.2
I. stationary	65.7	13.0	17.8	16.5	22.3	4.4	17.5	6.8
I. linear	15.2	31.9	6.3	49.7	12.9	10.4	20.1	39.8
I. quadratic	50.8	10.9	99.3	6.9	22.9	9.0	13.9	34.9
I. cubic	46.7	12.8	35.3	52.3	7.3	36.9	3.7	21.5
Ensemble	13.4	21.0	4.1	6.6	8.0	5.2	2.4	14.3

trends of occupancies change in the evolution of the phenomenon. Therefore, these CVs can be employed to detect time location of change points.

Acknowledgements. This research was co-funded by the Italian Complementary National Plan PNC-I.1 “Research initiatives for innovative technologies and pathways in the health and welfare sector” D.D. 931 of 06/06/2022, “DARE - DigitAl lifelong pRevEntion” initiative, code PNC0000002, CUP: B53C22006450001.

References

1. Durrleman, S., Richard, S.: Flexible regression models with cubic splines. *Stat. Med.* **8**, 551–561 (1989). <https://doi.org/10.1002/sim.4780080504>

2. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., Lovison, G.: An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biom. J.* **63**, 503–513 (2021). <https://doi.org/10.1002/bimj.202000189>

3. Fokianos, K.: Some recent progress in count time series. *Statistics* **45**, 49–58 (2011). <https://doi.org/10.1080/02331888.2010.541250>

4. Fokianos, K., Rahbek, A., Tjostheim, D.: Poisson autoregression. *J. Am. Stat. Assoc.* **104**, 1430–1439 (2009). <https://doi.org/10.1198/jasa.2009.tm08270>

5. Hastie, T., Tibshirani, R.: Generalized additive models: some applications. *J. Am. Stat. Assoc.* **82**, 371–386 (2014). <https://doi.org/10.1080/01621459.1987.10478440>

6. Miglio, M., Roli, G., Scagliarini, M.: A joint use of monitoring and forecasting methods to detect change points in daily hospitalizations. In: Programme and Abstracts, 24th International Conference on Computational Statistics, COMPSTAT 2022, pp. 22–23 (2022). <http://www.compstat2022.org>

7. Scott, S., Varian, H.: Predicting the present with Bayesian structural time series. *Int. J. Math. Model. Numer. Optim.* **5**, 4 (2014). <https://doi.org/10.1504/IJMMNO.2014.059942>



Bayesian Inference for Compartmental Models with Intractable Likelihood Function

Cecilia Viscardi¹(✉), Alessio Lachi^{1,2}, and Michela Baccini¹

¹ University of Florence, Department of Statistics, Computer Science, and Applications “Giuseppe Parenti” (DiSIA), Firenze, Italy
cecilia.viscardi@unifi.it

² Institute of Clinical Physiology of the Italian National Research Council (IFC-CNR), Epidemiology and Health Research, Pisa, Italy

Abstract. Compartmental models are widespread methods useful to understand and predict the dynamic of a phenomenon of interest, such as the spreading process of an infectious disease. Their formalization relies on systems of ordinary differential equations parameterized through some quantities of interest, such as the basic reproduction number and the duration of the infectious period. However, to account for the intrinsic uncertainty of an epidemic, we need a stochastic version of the equations based on a suitable statistical model. Unfortunately, the associated likelihood function is often intractable and requires specific methods for making inferences. In this work, we investigate three different strategies for addressing the intractability of the likelihood within a Bayesian framework. In particular, we consider Data-Augmentation and Pseudo-Marginal Markov Chain Monte Carlo algorithms, along with a likelihood-free approach based on Approximate Bayesian Computation. We describe and compare the methods at work on a simple SIR model to highlight their strengths and weaknesses.

Keywords: Compartmental models · Bayesian inference · Monte Carlo methods · Intractable likelihood

1 Introduction

Compartmental models are a class of models used to understand, describe, and predict the dynamic of a phenomenon of interest in a given population. Their mechanistic nature is well-suited for modeling infectious diseases. The main assumption of compartmental models is that the population of interest is divided, at each point in time, into compartments—i.e. non-overlapping groups of individuals that are homogeneous for some characteristic such as the health status. Starting from an initial composition of the population, transitions between compartments are allowed according to infectious dynamics described by a system of ordinary differential equations (ODEs). The system is governed by a vector of parameters that tune the transition rates. The ODEs solution is the deterministic evolution of the size of each compartment over time. However, a stochastic

extension of the system is essential to describe and quantify the uncertainty of the phenomenon. This extension relies on the assumption that transitions between compartments follow probabilistic rules defined by a statistical model. However, the likelihood function associated with the resulting stochastic model is often intractable or computationally demanding to evaluate, and specific inference methods are needed.

This paper aims to investigate the applicability of three different Bayesian estimation strategies. In particular, we want to provide a description and a comparison of their different ways of addressing the problem of the intractability of the likelihood. We consider a Data-Augmentation Markov Chain Monte Carlo (DA-MCMC) algorithm, a Pseudo Marginal Metropolis Hastings (PMMH) algorithm, and a likelihood-free approach based on Approximate Bayesian Computation (ABC). Finally, we test the methods at work on a simple SIR model, to assess their reliability and emphasize their strengths and weaknesses.

2 A Simple Example: The SIR Model

Let us consider the well-known Susceptible-Infected-Removed (SIR) model [5]. It considers three compartments, each one corresponding to a health status related to an infectious disease: S is the group of susceptible individuals who can potentially contract the disease when they come in contact with infectious individuals; I is the group of infectious individuals; R is the group of the so-called removed individuals who, after the infection, have become immune from the disease or have died.

The following system of difference equations is the discrete-time version of the well-known system of ODEs in [3] that describes the dynamic of the model:

$$\begin{cases} S(t) = S(t-1) - \pi_{SI}(t-1)S(t-1) \\ I(t) = I(t-1) + \pi_{SI}(t-1)S(t-1) - \pi_{IR}I(t-1) \\ R(t) = R(t-1) + \pi_{IR}I(t-1). \end{cases} \quad (1)$$

In Eq. (1), $\pi_{SI}(t-1)$ and π_{IR} are the (time-varying) probability of being infected and the (constant in time) probability of recovering/dying during a unit time interval, respectively. We assume that the waiting times before experimenting with an event have an exponential distribution, hence $\pi_{SI}(t-1) = 1 - \exp\left(-\frac{R_0}{\tau} \frac{I(t-1)}{S(0)}\right)$ and $\pi_{IR} = 1 - \exp(-\frac{1}{\tau})$, where τ is the average infectious period and R_0 is the basic reproduction number—i.e. the expected number of secondary infections caused by a single infected individual at the beginning of the epidemic.

To formalize a statistical model, we assume that, at each point in time, the number of new infections and new resolutions are random variables. Let $i_{t_1:t_2} = (i(t_1), \dots, i(t_2))$ and $r_{t_1:t_2} = (r(t_1), \dots, r(t_2))$ denote the random vectors of new infections and new resolutions from t_1 until t_2 (t_1 and $t_2 \in \{0, \dots, T\}$, $t_2 \geq t_1$), and by $i_{t_1:t_2}^*$ and $r_{t_1:t_2}^*$ their realization. Under the assumption that $i(t) \sim \text{Binom}(S(t), \pi_{SI}(t))$ and $r(t) \sim \text{Binom}(R(t), \pi_{IR})$ for each $t \in \{0, \dots, T\}$, we can write down a simple analytical form for the likelihood function based

on complete data—i.e. $i_{t_1:t_2}^*$ and $r_{t_1:t_2}^*$. However, when dealing with infectious diseases, there is often a problem of missing data: usually, we only observe the daily number of new infections, $i_{1:T}^*$, since recoveries are not accurately notified. If $r_{1:T}$ is a latent variable, the evaluation of the likelihood requires the following marginalization:

$$\begin{aligned} L(\theta \mid i_{1:T} = i_{1:T}^*) &= \sum_{r_{1:T}^* \in \mathcal{R}} \Pr(i_{1:T} = i_{1:T}^*, r_{1:T} = r_{1:T}^* \mid \theta) \\ &= \sum_{r_{1:T}^* \in \mathcal{R}} \prod_{t=1}^T \binom{S(t)}{i^*(t)} \pi_{\text{SI}}(t)^{i^*(t)} [1 - \pi_{\text{SI}}(t)]^{S^*(t) - i^*(t)} \binom{I^*(t)}{r^*(t)} \pi_{\text{IR}}^{r^*(t)} [1 - \pi_{\text{IR}}]^{I^*(t) - r^*(t)}. \end{aligned} \quad (2)$$

In Eq. (2), $S^*(t) = S^*(t-1) - i^*(t-1)$, $I^*(t) = I^*(t-1) + i^*(t-1) - r^*(t-1)$, and $(S(0), I(0), R(0))$, the initial condition of the system, is assumed to be known. Note that $\mathcal{R}$ is the space of all possible sequences $r_{1:T}^*$ —i.e. realizations of $r_{1:T}$ —that are compatible with the observed series $i_{1:T}^*$. The structure and the cardinality of $\mathcal{R}$ often make the point-wise evaluation of the likelihood computationally intensive or infeasible.

3 Three Different Approaches for Bayesian Inference

In Bayesian statistics, the parameters $\theta = (\tau, R_0)$ that govern the dynamic of the epidemic are considered as random variables with a prior distribution defined over the space Θ , here denoted by $\pi(\cdot)$. Thus, given the observed data $i_{1:T}^*$, our target is the posterior distribution derived through Bayes's formula:

$$\pi(\theta \mid i_{1:T}^*) \propto \pi(\theta) L(\theta \mid i_{1:T}^*) = \pi(\theta) \sum_{r_{1:T}^* \in \mathcal{R}} \Pr(i_{1:T} = i_{1:T}^*, r_{1:T} = r_{1:T}^* \mid \theta). \quad (3)$$

From Eq. (3) it is apparent that the computation of the posterior probabilities requires multiple evaluations of the intractable likelihood. We propose and compare three ways of overcoming this problem to get full Bayesian estimates.

3.1 Data Augmentation Metropolis-Hastings

A possible solution is that of getting samples from the posterior distribution $\pi(\theta, r_{1:T} \mid i_{1:T}^*)$ defined on the augmented space $\Theta \times \mathcal{R}$, rather than from $\pi(\theta \mid i_{1:T}^*)$. To this aim, Data Augmentation Markov Chain Monte Carlo methods (DA-MCMC) [6] can be implemented using the Metropolis-Hastings (MH) scheme displayed in Algorithm 1. Here, the key idea is that getting samples from the joint posterior avoids the use of the intractable likelihood since, at each iteration, we impute the missing series and compute an acceptance ratio based only on the complete likelihood $L(\theta \mid i_{1:T}^*, r_{1:T}^*)$. Since the produced Markov chain, has limiting distribution $\pi(\theta, r_{1:T} \mid i_{1:T}^*)$, after the assessment of the convergence, samples $\theta^{(0)}, \dots, \theta^{(S)}$ can be considered as drawn from the target $\pi(\theta \mid i_{1:T}^*)$.

Algorithm 1. Metropolis-Hastings

```

1: Initialize  $\theta^{(0)}, r_{1:T}^{(0)}$ 
2: for  $s$  in  $1 : S$  do
3:   Propose missing data  $r_{1:T}^* \sim q_r(\cdot \mid r_{1:T}^{(s-1)})$ 
4:   Propose  $\theta^* \sim q_\theta(\cdot \mid \theta^{(s-1)})$ 
5:   Compute  $\alpha = \min \left\{ 1, \frac{\pi(\theta^*)L(\theta^* \mid r_{1:T}^*, i_{1:T}^*)q_r(r_{1:T}^{(s-1)} \mid r_{1:T}^*)q_\theta(\theta^{(s-1)} \mid \theta^*)}{\pi(\theta^{(s-1)})L(\theta^{(s-1)} \mid r_{1:T}^{(s-1)}, i_{1:T}^*)q_r(r_{1:T}^* \mid r_{1:T}^{(s-1)})q_\theta(\theta^* \mid \theta^{(s-1)})} \right\}$ 
6:   Sample  $u \sim U(0, 1)$ 
7:   if  $u < \alpha$  then
8:     Set  $\theta^{(s)} = \theta^*$  and  $r_{1:T}^{(s)} = r_{1:T}^*$ 
9:   else
10:    set  $\theta^{(s)} = \theta^{(s-1)}$  and  $r_{1:T}^{(s)} = r_{1:T}^{(s-1)}$ 
11:   end if
12: end for

```

3.2 Pseudo Marginal Metropolis-Hastings

Another possible strategy is based on the formalization of the SIR model as a State Space Model (SSM). This approach allows the use of a Pseudo Marginal Metropolis-Hastings algorithm (PMMH) to get samples directly from the *marginal* posterior, $\pi(\theta \mid i_{1:T}^*)$ (see [2]). It relies on an *approximation* of the likelihood obtained by running a Particle Filter (PF) algorithm at each iteration of the MH. To formalize our model as SSM, without introducing assumptions different from those used in the other two methods, we propose to consider the two following discrete-time processes: 1) the latent process $X_{1:T} = (S(1), R(1)), \dots, (S(T), R(T))$; 2) the observed process $Y_{1:T} = i_{1:T}$. According to the probabilistic rules defined in Sect 2, we can show that 1) $\Pr(X_t = x_t \mid x_{t-1}) = \text{Bin}(i^*(t-1); \pi_{\text{SI}}(t-1), S(t-1)) \cdot \text{Bin}(r^*(t-1); \pi_{\text{IR}}, I(t-1))$; 2) $\Pr(Y_t = y_t \mid x_{t-1}) = \Pr(i(t) = i^*(t) \mid x_t) = \text{Bin}(i^*(t); \pi_{\text{SI}}(t), S(t))$.

This allows us to estimate $L(\theta \mid i_{1:T}^*) = \Pr(Y_{1:T} = i_{1:T}^*)$ using a PF and use it to replace the intractable likelihood in Algorithm 2.

Algorithm 2. Pseudo-Marginal Metropolis-Hastings

```

1: Initialize  $\theta^{(0)}$ 
2: for  $s$  in  $1 : S$  do
3:   Propose  $\theta^* \sim q_\theta(\cdot \mid \theta^{(s-1)})$ 
4:   Compute the estimate of the likelihood function  $\hat{L}(\theta^* \mid i_{1:T}^*)$  running a PF
5:   Compute  $\alpha = \min \left\{ 1, \frac{\pi(\theta^*)\hat{L}(\theta^* \mid i_{1:T}^*)q_\theta(\theta^{(s-1)} \mid \theta^*)}{\pi(\theta^{(s-1)})\hat{L}(\theta^{(s-1)} \mid i_{1:T}^*)q_\theta(\theta^* \mid \theta^{(s-1)})} \right\}$ 
6:   Sample  $u \sim U(0, 1)$ 
7:   if  $u < \alpha$  then
8:     Set  $\theta^{(s)} = \theta^*$ 
9:   else
10:    set  $\theta^{(s)} = \theta^{(s-1)}$ 
11:   end if
12: end for

```

3.3 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a likelihood-free method that dispenses with likelihood computation and only requires the availability of a computer program, usually called “simulator”, that reproduces the stochastic data generative process [7]. The underlying idea is that samples from the prior distribution can be converted into samples from the posterior through three simple steps: 1) draw S parameter proposals from $\pi(\cdot)$; 2) give each parameter proposal, $\theta^{(s)}$, as an input to the simulator $\text{Pr}(\cdot \mid \cdot)$ to produce pseudo-data $\widehat{i}_{1:T}(\theta^{(s)})$; 3) retain only parameter proposals such that $\rho(\widehat{i}_{1:T}(\theta^{(s)}); i_{1:T}^*) < e$, where $\rho(\cdot; \cdot)$ is a suitable distance function and e is a positive tolerance threshold.

More sophisticated sampling schemes are based on a decreasing sequence of thresholds, rather than a fixed tuning parameter. Examples are the Population Monte Carlo ABC [1] and some adaptive versions inspired by it [4], such as the one summarized in Algorithm 3.

Algorithm 3. Adaptive Population Monte Carlo ABC

```

1: Initialize  $e_1$ 
2: for  $j$  in  $1 : M$  do
3:   Simulate  $\theta_j^{(1)} \sim \pi(\cdot)$  and  $\widehat{i}_{1:T}(\theta_j^{(1)}) \sim \text{Pr}(\cdot \mid \theta_j^{(1)})$  until  $\rho(\widehat{i}_{1:T}(\theta_j^{(1)}); i_{1:T}^*) < e_1$ 
4:   Set  $\omega_j^{(1)} = \frac{1}{M}$ 
5: end for
6: Select  $e_2$  using an adaptive strategy.
7: for  $s$  in  $2 : S$  do
8:   Set  $\Sigma_s$  to twice the empirical covariance matrix of  $\theta_1^{(s-1)}, \dots, \theta_M^{(s-1)}$ 
9:   for  $j$  in  $1 : M$  do
10:    Pick  $\theta_j^*$  from  $(\theta_1^{(s-1)}, \dots, \theta_M^{(s-1)})$  with probabilities  $(\omega_1^{(s-1)}, \dots, \omega_M^{(s-1)})$ 
11:    Generate  $\theta_j^{(s)} \mid \theta_j^* \sim \text{MVN}(\theta_j^*, \Sigma_s)$  and  $\widehat{i}_{1:T}(\theta_j^{(s)}) \sim \text{Pr}(\cdot \mid \theta_j^{(s)})$ 
12:    Set  $\omega_j^{(s)} \propto \frac{\pi(\theta_j^{(s)})}{\sum_{m=1}^M \omega_m^{(s-1)} \phi\{\tau_s^{-1}(\theta_m^{(s)} - \theta_m^{(s-1)})\}}$   $\mathbf{1}\{\rho(\widehat{i}_{1:T}(\theta_j^{(s)}); i_{1:T}^*) < e_s\}$ 
13:    where  $\phi$  represents the density of the Standard Normal distribution
14:   end for
15:   Select  $e_{s+1}$  using an adaptive strategy.
16: end for
    
```

The output is a sample from a target distribution different from that of the previous methods since it is an *approximate* posterior distribution:

$$\pi_{e_S}(\theta \mid i_{1:T}^*) \propto \pi(\theta) \text{Pr}(\rho(\widehat{i}_{1:T}(\theta^{(s)}); i_{1:T}^*) < e_S).$$

Notice that, under mild conditions, $\text{Pr}(\rho(\widehat{i}_{1:T}(\theta^{(s)}); i_{1:T}^*) < e_S) = L(\theta \mid i_{1:T}^*)$ when $e_S = 0$. Thus, the greater the value of e_S , the greater the approximation w.r.t. the true posterior.

4 Results and Discussion

For the sake of comparison among methods, we considered as observed data $i_{1:90}^*$ the output of a simulation from the model with parameters $R_0^{\text{true}} = 3$ and $\tau^{\text{true}} =$

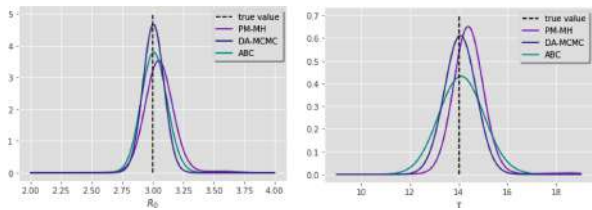


Fig. 1. Comparison among posterior distributions for R_0 (left) and τ (right).

Table 1. Bayesian Mean Square Error (BMSE)

	R_0	τ
DA-MCMC	0.0037	0.2162
PMMH	0.0114	0.4748
ABC	0.0057	0.4385

14. We set uniform prior distributions: $R_0 \sim \text{Unif}(0, 6)$ and $\tau \sim \text{Unif}(7, 14)$. The proposal distribution $q_r(\cdot \mid \cdot)$ in Algorithm 1 is that suggested in [6], while $q_\theta(\cdot \mid \cdot)$ are Normal distributions. We gave all the algorithms a budget of time of 10 min. Looking at Fig. 1, we can see that their outputs are quite similar in terms of approximation of the posterior distributions. This is a piece of evidence in favor of the reliability of all the methods. To compare the ability to retrieve the “true” values, we computed the Bayesian Mean Squared Errors. Looking at Table 1 we can see that DA-MCMC outperforms the other methods. However, since it defines the chain on an augmented space, the produced Markov chain is strongly autocorrelated and a tremendous number of iterations are required to provide a valid sample. We speculate that this problem is more serious when the compartmental model is more complex than the SIR model, since the size of the latent variable space could be greater. PMMH does not suffer from this problem but it is highly sensitive to the quality of the likelihood approximation provided by PF, and this approximation could be particularly challenging in more complex models. Finally, ABC, while prone to an approximation error induced by the threshold ϵ_S , is highly parallelizable and seems to be more promising in terms of extensions to complex cases.

References

1. Beaumont, M., Cornuet, J., Marin, J., Robert, C.: Adaptive approximate Bayesian computation. *Biometrika* **96**(4), 983–990 (2009)
2. Chopin, N., Papaspiliopoulos, O.: An Introduction to Sequential Monte Carlo, vol. 4. Springer (2020)
3. Hethcote, H.W.: The mathematics of infectious diseases. *SIAM Rev.* **42**(4), 599–653 (2000)
4. Lenormand, M., Jabot, F., Deffuant, G.: Adaptive approximate Bayesian computation for complex models. *Comput. Statist.* **28**(6), 2777–2796 (2013)
5. Mckendrick, A.: Applications of mathematics to medical problems. *Proc. Edinb. Math. Soc.* **44**, 98–130 (1925)
6. O’Neill, P., Roberts, G.: Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A (Statist. Soc.)*, **162**(1), 121–129 (1999). Publisher: [Wiley, Royal Statistical Society]
7. Sisson, S., Fan, Y., Beaumont, M. (eds.): Handbook of Approximate Bayesian Computation. CRC Press, Taylor & Francis Group, Boca Raton (2019)

Author Index

A

Adamo, Domenico 282
Adelfio, Giada 588
Agnes, Annamaria 501
Ambrosetti, Elena 33
Antonelli, Barbara 240
Antonelli, Manuela 630
Antonucci, Laura 1
Arboretti, Rosa 276
Arcaio, Micaela 398
Arezzo, Maria Felice 618
Aria, Massimo 126
Arias-Salazar, Alejandra 300
Arnone, Eleonora 361, 386
Ascari, Roberto 96
Augugliaro, Luigi 58

B

Baccini, Michela 655
Bagnasco, Giulia 398
Ballante, Elena 455
Ballerini, Veronica 160
Banerjee, Sudipto 442
Bartolucci, Francesco 581
Bazzurri, Giorgia 330
Belmont, Jafet 379
Bernuy, Victoria San Juan 223
Berta, Giulia 251
Bertarelli, Gaia 563
Bertoni, Marco 77
Bianchi, Gianpiero 282
Bianchino, Antonella 410, 416
Biondi, Alberto 501
Bisaglia, Luisa 90
Bombelli, Ilaria 540
Botte, Brunella 306
Bovera, Filippo 194
Brentari, Eugenio 552
Brizio, Cesare 479

C

Campostrini, Stefano 563
Canale, Antonio 337
Cantone, Giulio Giacomo 601
Cantoni, Beatrice 630
Cappozzo, Andrea 240
Carbonara, Monica 410
Carey, Michelle 166
Carucci, Agata Maria Madia 416
Casacchia, Oliviero 373
Chan, Lax 218
Chen, Yunxiao 263
Chokami, Amir Khorrami 288
Ciano, Tiziana 324
Clemente, Aldo 386
Clementi, Letizia 361
Colosimo, Bianca Maria 404
Conti, Cinzia 373
Conti, Pier Luigi 234
Coppellotti, Matteo 40
Coronnello, Claudia 58
Crisafulli, Saverio Gianluca 490
Crosino, Andrea 524
Cucinella, Morena 398

D

D'Angelo, Nicoletta 588
D'Ugo, Domenico 501
Damiano, Rodolfo 546
De Gregorio, Alessandro 46
de Magistris, Paolo Santucci 200
Dede', Luca 404
Del Rey, A. 342
Di Carlo, Luigi 479
Di Fonzo, Tommaso 119
Dietrich, Hans 33
Disegna, Marta 276
Doretti, Marco 142

E

Epifani, Ilenia 630

F

Fabrizi, Enrico 246
 Fabrizio, Paolomaria 392
 Fellingham, Gilbert W. 318
 Ferrante, Maria Rosaria 506
 Ferrara, Massimiliano 324
 Ferraro, Maria Brigida 28
 Ferré, Francesca 608
 Figini, Silvia 455
 Fioredistella Iezzi, Domenica 540
 Firza, Najada 1
 Fokkema, Marjolein 126
 Fontana, Roberto 9
 Forciniti, Alessia 15, 495
 Forlani, Marco 240
 Foroni, Beatrice 22
 Forti, Marco 28
 Fortunato, Cecilia 33
 Frigeri, Michela 40
 Frisardi, Dario 46
 Frumento, Paolo 246
 Furlan, Claudia 52

G

Gagliano, Antonino 58
 Gallo, Alessandro 64
 García-Gómez, J. 342
 Gardini, Aldo 71
 Gauglitz, Günter 404
 Genesin, Leonardo 77
 Genre, Andrea 524
 Gentile, Susanna 83
 Gerolimetto, Margherita 90
 Giacalone, Massimiliano 429, 490
 Giampino, Alice 96
 Gianella, Matteo 102
 Gilardi, Andrea 108
 Gilioli, Gianni 257
 Giordani, Paolo 28, 113
 Giordano, Sabrina 643
 Girolimetto, Daniele 119
 Gnasso, Agostino 126
 Golia, Silvia 130
 Gottard, Anna 136, 643
 Gourgoura, Konstantina 142

Greco, Fedele 71, 148
 Greco, Luca 355
 Grossi, Giulio 154, 160
 Guagnano, Giuseppina 618
 Guastadisegni, Lucia 148
 Guerra, Marco 9
 Guetto, Raffaele 392
 Guglielmi, Alessandra 40, 102

H

Higgins, Catherine 166
 Höhne, Jan Karem 483

I

Iacus, Stefano M. 46
 Iafrate, Francesco 46
 Iaquinta, Massimo 410
 Ieva, Francesca 108, 240
 Ievoli, Riccardo 171
 Illian, Janine 379
 Iorio, Carmela 126
 Ippoliti, Luigi 479

J

Jacques, Julien 637
 Javed, Aamir 177

K

Khan, Feroz 183
 Khorrami Chokami, Amir 188
 Koechlin, Guillaume 194
 Krauth, Alexander 404

L

La Grutta, Stefaina 366
 Lacava, Demetrio 200
 Lachi, Alessio 655
 Lanfiuti Baldi, Giacomo 206
 Lazzara, Valentina 366
 Leombruni, Roberto 212
 Leoni, Olivia 240
 Levantesi, Susanna 113
 Lideikyte-Huber, Giedre 435
 Locatelli, Elena 501
 Lodetti, Patrizio 223, 392
 Loffredo, Giuseppe 230
 Longobardi, Maria 429

Lorenzon, Laura 501
Luta, George 355

M

M. Sangalli, Laura 361
Malizia, Velia 366
Marchesin, Leonardo 40
Marella, Daniela 234
Maria Quetti, Federico 455
Marinensi, Giada 306
Martino, Sara 379
Masci, Chiara 240, 461
Maturo, Fabrizio 230
Mazzitelli, Dante 1
Meggiolaro, Silvia 52, 613
Menafoglio, Alessandra 404
Menardi, Giovanna 77
Merlo, Luca 22, 246
Mezzetti, Maura 330
Miaci, Eleonora 251, 449
Miglio, Rossella 649
Migliorati, Manlio 257
Migliorati, Sonia 96
Mignemi, Giuseppe 263
Mimmo, Angelo 269
Molena, Alberto 276
Mongelli, Lucia 282
Montagna, Silvia 188, 288
Morabito, Maria Francesca 294
Moretti, Angelo 300
Moustaki, Irini 263
Muggeo, Vito 366
Musella, Flaminia 306
Mussini, Mauro 312

N

Narcisi, Martina 318
Nava, Consuelo Rubina 324
Negri, Ilia 330
Nicolussi, Federica 404
Nigri, Andrea 113, 206
Novelli, Marco 511

O

Oddo, Giuseppe 490
Ongaro, Fausta 613
Onorati, Paolo 337
Orfao, G. 342
Otranto, Edoardo 349

P

Pacillo, Simona 355
Paganoni, Anna Maria 240, 461
Page, Garritt L. 318
Paggiaro, Adriano 77
Palazzo, Lucio 171
Palladino, Benedetta E. 294
Palumbo, Roberto A. 410
Palummo, Alessandro 361
Pampillonia, Santina 398
Pandolfo, Alessandra 366
Pangallo, Camilla 373
Panichella, Nazareno 251, 449
Panunzi, Greta 379
Panzeri, Agnese 136
Panzeri, Simone 386
Parrado, E. 342
Parroco, Anna Maria 398
Patanè, Giulia 404
Pellicani, Michela C. 410
Peluso, Alfonso 416
Persiani, Roberto 501
Petrella, Lea 22, 246
Piccitto, Giorgio 422, 426
Piscitelli, Alfonso 552
Piscopo, Gianfranco 429, 490
Pittavino, Marta 435
Policastro, Valeria 495
Presicce, Luca 442
Priulla, Andrea 449

R

Ragni, Alessandra 461
Ragozini, Giancarlo 171, 495
Rapposelli, Agnese 177, 183
Rastelli, Riccardo 581
Ricciotti, Lorena 467
Rizzelli, Stefano 473
Rocco, Emilia 154
Rojas-Perilla, Natalia 300
Roli, Giulia 649
Romagnoli, Luca 479
Romano, Elvira 230
Rondinelli, Roberto 171
Rotondo, Antonella 410
Rottino, Fabio Massimo 373
Rubina Nava, Consuelo 212

S

Salvati, Nicola 246
 Salvatore, Camilla 483
 Sambucini, Valeria 83
 Sangalli, Laura M. 386
 Santarcangelo, Vito 490
 Santelli, Francesco 15, 495
 Santoro, Gloria 501
 Scaffidi Domianello, Luca 349
 Scagliarini, Michele 649
 Scarpa, Silvia 506
 Scarso, Leonardo 511
 Schiavon, Lorenzo 563
 Schoier, Gabriella 429
 Sciandra, Andrea 517
 Sciarabba, Martina 251
 Sciaraffa, Nicolina 58
 Sciascia, Ivan 524
 Scisci, Domingo 426
 Scricciolo, Catia 528
 Scrucca, Luca 534
 Secchi, Piercesare 194
 Seghieri, Chiara 608
 Seri, Emiliano 540
 Shatarat, Omar 366
 Silverman, Bernard W. 218
 Simonetti, Andrea 546
 Simonetto, Anna 257, 552
 Sottile, Gianluca 58, 366
 Sperlich, Stefan 506
 Stanghellini, Elena 142
 Stefanizzi, Sonia 426
 Stefanucci, Marco 556
 Stival, Mattia 563
 Stolf, Federica 569
 Striessnig, Erich 624

T

Tafuro, Sara 574
 Tancini, Daniele 581
 Tarantino, Marco 588
 Tedesco, Lorenzo 595
 Tirelli, Flavio 501
 Tocchioni, Valentina 294
 Tokdar, Surya T. 288
 Tomaselli, Venera 601
 Toninelli, Daniele 269
 Tortú, Costanza 608
 Tosi, Marco 613
 Trappolini, Eleonora 618
 Trimarchi, Alessandra 624
 Trivisano, Carlo 71, 148
 Trovò, Francesco 630
 Tuzzi, Arjuna 517

U

Ursino, Bruno 630

V

Ventura, Matteo 637
 Verbena, Serena 294
 Verrina, Vanessa 643
 Vesely, Anna 649
 Vicari, Donatella 113
 Viegì, Giovanni 366
 Vignoli, Daniele 64, 392
 Violante, Francesco Saverio 511
 Viscardi, Cecilia 655

Z

Zanatta, Marina 212
 Zuccolotto, Paola 637