Italian Statistical Society Series on Advances in Statistics

Alessio Pollice · Paolo Mariani *Editors*

Methodological and Applied Statistics and Demography III

SIS 2024, Short Papers, Contributed Sessions 1

SIS Società Italiana di Statistica

Springer

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 III

SIS 2024, Short Papers, Contributed Sessions 1

Springer

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 978-3-031-64430-6 ISBN 978-3-031-64431-3 (eBook)
https://doi.org/10.1007/978-3-031-64431-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025
Chapter "A Longitudinal Study of the Gender Gap in School Grades via Flexible Bayesian Beta Regression" is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). For further details see license information in the chapter.

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 go 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 third volume presents a selection of contributions from authors who delivered presentations during the Contributed sessions.

June 2024

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

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

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

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

Contributed Sessions I

Bayesian Inference for Generalized Linear Models via Quasi-Posteriors: an Application to Eurasian Chaffinch Abundance in Finland 3
Davide Agnoletto, Tommaso Rigon, and David B. Dunson

A Spatial Product Partition Model for PM_{10} Data 8
Luca Aiello, Sirio Legramanti, and Lucia Paci

Efficient Bayesian Estimation of Spatial Poisson Auto-regression with Leroux Random Effects 14
Pierfrancesco Alaimo Di Loro, Dankmar Böhning, and Sujit K. Sahu

A Two-Step Approach to Analyze the Inclusivity Distribution in Childcare Services in Italy 20
Angela Andreella, Gaia Bertarelli, Federico Caldura, and Stefano Campostrini

Analyzing Compression of Comorbidity in Italy: A Mixture Model with Regional Random Effects 27
Angela Andreella and Stefano Campostrini

Beyond Correlation: Discovering Causal Effects of AI on Innovation Across European SMEs 34
Lea Anna Cozzucoli and Francesco Pauli

Simultaneous Clustering and Dimension Reduction of Functional Data 40
Roberto Rocci and S. Antonio Gattone

Comparing Methods for Testing Equality of Means in One-Way Designs 46
Rosa Arboretti, Elena Barzizza, and Nicolò Biasetton

Evaluating the Geographical Patterns of Intimate Partner Violence in Nigeria, Zambia, and Zimbabwe Using INLA-SPDE Approach 53
Micaela Arcaio, Anna Maria Parroco, and Chibuzor Christopher Nnanatu

A Two-Part Beta Regression Model with Measurement Error 58
Maria Felice Arezzo, Giuseppina Guagnano, and Domenico Vitale

TALL: A New Shiny App for Text Analysis 64
Massimo Aria, Corrado Cuccurullo, Luca D'Aniello, Michelangelo Misuraca, and Maria Spano

On the Estimation of INAR Models with Pseudo-Variance Quasi-Maximum Likelihood 71
Mirko Armillotta and Paolo Gorgi

Estimating Spatial Anisotropy in Semiparametric Regression with Differential Regularization 77
Eleonora Arnone, Matteo Tomasetto, and Laura M Sangalli

A New Distribution for Overdispersed Count Data 82
Roberto Ascari, Agnese M. Di Brisco, and Sonia Migliorati

Uncertainty Reduction in a Class of Dependent Dirichlet Processes 88
Filippo Ascolani, Stefano Damato, and Matteo Ruggiero

Non-Informative Priors in Wright–Fisher Smoothing 93
Filippo Ascolani, Ylenia F. Buttigliero, and Matteo Ruggiero

Parents or Peers? Generational Socialisation and Attitudes Towards Migrants Across European Societies 98
Leo Azzollini, Daniela Bellani, and Giulia Rivellini

Variable Selection in Multilevel Models via Knockoffs: the Case of INVALSI National Tests 104
Silvia Bacci, Emanuela Dreassi, Leonardo Grilli, and Carla Rampichini

A Fully Rank-Based Hierarchical Clustering Approach Using Information Imbalance 109
Elena Ballante, Antonietta Mira, and Silvia Figini

Non-parametric Similarity Measure Based on Information Imbalance 115
Elena Ballante, Silvia Figini, and Antonietta Mira

Data Transformation and Its Validity in a Two-Sample Problem: An Illustration Based on Graphical Models 121
Erika Banzato, Davide Risso, Monica Chiogna, and Vera Djordjilović

Object-Oriented Analysis of Spatial Complex Data in Stream-Network Domains 127
Chiara Barbi, Alessandra Menafoglio, and Piercesare Secchi

Latent Class Multi-state Quantile Regression 133
Rosario Barone and Alessio Farcomeni

ChatGPT and Conversational AI in Customer Satisfaction: A Preliminary Literature Review 139
Elena Barzizza, Nicolò Biasetton, and Luigi Salmaso

Eco-Sustainable Tourism in Europe, an Analysis with Multilevel Models 145
Francesca Bassi and Juan Antonio Marmolejo Martìn

Extending the Vertical Model: An Alternative Approach to Competing Risks with Clustered Data 152
Salvatore Battaglia, Marta Fiocco, and Hein Putter

Likelihood-Based Boosting for Variance Components Selection in Linear Mixed Models 158
Michela Battauz and Paolo Vidoni

Financial Institution Confidence Index with Natural Language Processing 164
Leonardo Becarelli and Luca Trapin

Nonparametric Space-Time Density Estimation for Point Processes over Irregular Regions 170
Blerta Begu, Simone Panzeri, Eleonora Arnone, Michelle Carey, and Laura M. Sangalli

Generalized Additive Mixed Models in Medicine: A Case Study on LDL Cholesterol in People Living with HIV Under Different Antiretroviral Regimens 176
Michele Bellomo, Nicola Gianotti, Camilla Muccini, Hamid Hasson, Diana Canetti, Silvia Nozza, Monica Guffanti, Emanuela Messina, Vincenzo Spagnuolo, Riccardo Lolatto, Sara Diotallevi, Antonella Castagna, and Laura Galli

An Age-Adjusted Disproportionality Measure for Pediatric Pharmacovigilance 182
Pietro Belloni and Federica Pizzolato

Country Prevalence and Problematic Gambling Among European Adolescents 189
Elisa Benedetti, Gabriele Lombardi, Rodolfo Cotichini, Sonia Cerrai, Marco Scalese, and Sabrina Molinaro

Stochastic Variational Inference for Structured Bayesian Hierarchical Models 196
Massimo Bilancia and Andrea Nigri

Advances in Intermittent Demand Forecasting 202
Parvaneh Rafieisangari and Luisa Bisaglia

Life Satisfaction and Institutional and Interpersonal Trust: A Comparative Analysis of Italian and Foreign Citizens 208
Martina Bodo, Angela Paparusso, and Elena Ambrosetti

Population Change Across Italian Municipalities: Urbanisation and the Role of Connectivity 215
Danilo Bolano, Daniela Ghio, Alessandro Di Nallo, and Francesco Billari

A Bayesian Two-Layer Latent Variable Model for Protein Inference 221
Jordy Bollon and Simone Tiberi

Assessing Variables Importance When Clustering Enterprises 226
Ilaria Bombelli, Alessio Guandalini, and Giorgia Sacco

A Longitudinal Study of the Gender Gap in School Grades via Flexible Bayesian Beta Regression 233
Laura Bondi, Beatrice Franzolini, and Marco Palma

Confidence Interval for the Complexity Index of Functional Data 239
Enea G. Bongiorno, Lax Chan, and Aldo Goia

Studying a Mixture of Functional Data Having Different Complexity 243
Enea G. Bongiorno, Lax Chan, and Aldo Goia

Partial M-Quantile Regression for Predictive Mantainance 248
Riccardo Borgoni, Enrico Fabrizi, Nicola Salvati, Francesco Schirripa Spagnolo, and Diego Zappa

Completion of Partially Observed Multivariate Functional Data 254
Marco Borriero, Luigi Augugliaro, Salvatore Latora, and Veronica Vinciotti

Approximate Inference for the Bayesian Discrepancy Measure for Precise Statistical Hypotheses 260
Elena Bortolato, Francesco Bertolino, Monica Musio, and Laura Ventura

A Latent Curve Model to Estimate the Evolution of Urban Air Pollution 266
Bottazzi Schenone Mariaelena and Grimaccia Elena

Measures of Interrater Agreement for Ratings on Nominal Scale 273
Giuseppe Bove and Daniela Marella

Median-Based Inequality Measures 279
Vytaras Brazauskas, Francesca Greselin, and Ričardas Zitikis

Improving Psychological Digital Wellbeing in Young Adult Users: A Bayesian Networks Approach 285
Chiara Brombin, Clelia Di Serio, and Paola Vicard

Greening Obligations, Avoidance Behavior, and Farms' Environmental and Economic Performance: Insights from FADN 291
Zelda Brutti, Marzia Freo, and Laura Serlenga

Trimodal Extension Based on the Flexible Generalized Skew-Normal Distribution 297
Michele Bufalo and Andrea Nigri

Assessing Literacy Proficiency of Students with a Migratory Background: The Case of Lombardia Region 303
Alessio Buonomo, Stefania Capecchi, Francesca Di Iorio, and Patrizia Giannantoni

Modeling and Synthesis of GPS Trajectories for Privacy-Preserving Mobility Data Generation 309
Arianna Burzacchi, Lise Bellanger, Klervi Le Gall, Aymeric Stamm, and Simone Vantini

A Modeling Approach to Investigate Species-Environment relationship in Coral Areas Using *in-Situ* observations 315
Crescenza Calculli, Diego Battagliese, Porzia Maiorano, Angela Carluccio, and Gianfranco D'Onghia

Local and Global Economic Policy Uncertainty Influence on US Stock Market Volatility 321
Vincenzo Candila, Oguzhan Cepni, Giampiero M. Gallo, and Rangan Gupta

An EWMA Control Chart for Multivariate Functional Data 328
Christian Capezza, Giovanna Capizzi, Fabio Centofanti, Antonio Lepore, and Biagio Palumbo

Low-Income Earners: New Possibilities of Study from Administrative Archives 334
Monica Carbonara and Sara Casacci

Inequalities Among Older People: Exploring Differences in Deprivation Patterns 341
Maria Carella, Roberta Misuraca, and Anna Paterno

The Multifaced Challenge to Tackle Energy Poverty: Outlook Scenarios for EU Countries 347
Alfonso Carfora and Giuseppe Scandurra

Bayesian Models for Registration of Functional Data 354
Jacopo Gardella, Rossella Carone, Simone Ciardulli, Nicola Francescon, Matteo Freddi, Elisa Garlanda, Alessandro Casa, and Alessia Pini

Regionalization Methods for Compound Extremes Based on the Wasserstein Distance 360
Regina Castrovilli, Fabrizio Durante, Daniela Gallo, and Gianfausto Salvadori

Efficient Parametric Tests in Semiparametric Regression with Differential Regularization 366
Michele Cavazzutti, Eleonora Arnone, Federico Ferraccioli, Livio Finos, and Laura M. Sangalli

Bispectral Analysis of Markov Switching Bilinear Models 373
Maddalena Cavicchioli, Ahmed Ghezal, and Imane Zemmouri

Topological Data Analysis Applied to Radiomics (topiomics) Data in Recurrent Prostate Cancer 379
Lara Cavinato, Lorenzo Ferrara, Matteo Pegoraro, Paola Anna Erba, and Francesca Ieva

PixHomology: A New Algorithm for Computing Persistent Homology for Pixel Data 386
Riccardo Ceccaroni and Pierpaolo Brutti

A Multi-Aspect Permutation Test for Selecting the Optimum Number of Clusters 393
Riccardo Ceccato, Elena Barzizza, Nicoló Biasetton, and Marta Disegna

Multilevel Bivariate Areal Modelling for School Data: An Application with R-INLA 399
Leonardo Cefalo and Virgilio Gómez-Rubio

RODESSA for Robust Multivariate Singular Spectrum Analysis 406
Fabio Centofanti, Mia Hubert, Biagio Palumbo, and Peter J. Rousseeuw

Sequence Analysis for Vaccination Data 412
Chiara Chiavenna, Danilo Bolano, and Alessia Melegaro

A Flexible Parametric Approach to Synthetic Patients Generation in Clinical Trials 419
Marta Cipriani, Lorenzo Di Rocco, and Marco Alfò

A Multi-Domain Model with Partial Differential Regularization: An Application to Neuroimaging Data 426
Aldo Clemente, Eleonora Arnone, and Laura M. Sangalli

Standardization Process of Budget Administrative Data, Having Statistic Characteristics and Usefulness for Public Finance Control 431
Daniela Collesi

The Italian Statistics Olympiads 438
Patrizia Collesi and Antonio Pitrone

Learning Block Structures in Gaussian Graphical Models for Spectrometric Data Analysis 444
Alessandro Colombi, Lucia Paci, and Alessia Pini

On the Evidence of a New Bayesian Discrepancy Measure for Discrete Scalar Parameters 450
Silvia Columbu, Francesco Bertolino, and Monica Musio

A Direct Estimate of Lorenz Curve and Inequality Measures in European Countries 457
Francesca Condino

A Kruskal-Wallis Type Test for Multi-treatment Effect 463
Pier Luigi Conti and Livia De Giovanni

A Bayesian Overlapping Stochastic Block Model for Clustering Biographical Networks 469
Noemi Corsini and Michael Fop

Exploring Sustainability Amidst Global Challenges: A Focus on Cooperative Enterprises 475
Michele Costa and Flavio Delbono

Goodness of Fit Measures for Generalized Estimating Equations. An Empirical Evidence on Relationships Between Gender Diversity and Corporate Performance 480
Anna Crisci, Luana Serino, Pasquale Sarnacchiaro, and Francesco Campanella

Educational Disparities in the Risk of Dementia and Subsequent Risk of Hospitalization in Lazio Region 485
Agostino Cristofalo, Silvia Cascini, Giulia Cesaroni, Eleonora Trappolini, Nera Agabiti, and Anna Maria Bargagli

Exploring Data Sharing Behaviour Using both Theory and Data-Driven Methods 491
Federica Cugnata, Chiara Brombin, Roberto Buccione, and Clelia Di Serio

Is Total Expenditure a Valid Proxy for Income in Statistical Matching? a Counterfactual Experiment 497
Elena Dalla Chiara, Marcello D'Orazio, and Federico Perali

Exact Sampling of Two Coagulated Partitions 503
Marco Dalla Pria, Matteo Ruggiero, and Dario Spanò

A New Competition Model Combining the Lotka-Volterra Model and the Bass Model in Pharmacological Market Competition 507
Alessandra Dalla Valle and Cinzia Mortarino

Measuring Sustainability in Tourism via a Multi Criteria Decision Making Approach 513
G. Damiana Costanzo, Barbara Dattilo, Mascia Di Torrice, Michelangelo Misuraca, and Maria Spano

Biological Species Delimitation Based on Genetic and Spatial Dissimilarity Using Partial Mantel Tests 519
Gabriele d'Angella and Christian Hennig

Smooth Logistic Mass Univariate Inference for MS Lesion Data Using Sign-Flipping 525
Samuel Davenport, Riccardo de Santis, Jesse Hemerik, Jelle Goeman, and Livio Finos

Penalised Spatial Quantile Regression: Application to Air Quality Data 532
Marco F. De Sanctis, Ilenia Di Battista, Eleonora Arnone, Cristian Castiglione, Mauro Bernardi, Alessandro Palummo, and Laura M. Sangalli

Sign-Flip Test for the Cox Regression Model 538
Riccardo De Santis, Jelle J. Goeman, Hein Putter, and Livio Finos

The Distribution of the Risk Function for Interval Estimation 543
Fulvio De Santis, Stefania Gubbiotti, and Francesco Mariani

Relational Hypervents Models for Co-authorship Networks: The Case of the Italian Academic Statisticians 549
Domenico De Stefano, Amin Gino Fabbrucci Barbagli, Francesco Santelli, and Susanna Zaccarin

Misogynistic Content Detection Over Social Networks 555
Emiliano del Gobbo, Elisa Ignazzi, Barbara Cafarelli, and Lara Fontanella

Evaluation of Adjustment for Unobserved Confounders in Occupational Epidemiology Studies 561
Silvia D'Elia, Stefania Massari, Marco Alfò, Claudio Gariazzo, Alessandro Gialluisi, Alessandro Marinaccio, Licia Iacoviello, and Simona Costanzo

Digitalization, Work-Related Risk Factors and Well-Being: Importance and Interactions from Tree-Based Methods 567
Fabio Demaria and Maddalena Cavicchioli

Individual Causation with Additional Information 574
Giuseppe Demuru, Monica Musio, and A. Philip Dawid

Indexing Social Unfitness: Quantifying Ordinal Levels by Problems Severity 580
Giulio D'Epifanio

A Markov-Switching Model for Studying Extreme Temperatures in *Friuli Venezia Giulia*, Italy 586
Gioia Di Credico, Vincenzo Gioia, and Francesco Pauli

Estimating Cure Fraction in Population-Based Survival Colon Cancer Data 592
Fabrizio Di Mari, Roberto Rocci, Giovanna Tagliabue, Silvia Rossi, and Roberta De Angelis

Lower Bias Circular Density Estimation with Contaminated Data 599
Marco Di Marzio, Stefania Fensore, Agnese Panzera, and Chiara Passamonti

A New Methodological Approach to Build a Multidimensional Inequality Composite Indicator 605
Cinzia Di Nuzzo, Venera Tomaselli, and Gianpiero Torrisi

Regularized Functional Partial Least Squares Regression of Neuroimaging Data 611
Pietro Donelli, Harold A. Hernández-Roig, M. Carmen Aguilera-Morillo, Eleonora Arnone, Letizia Clementi, Rosa E. Lillo, Alessandro Palummo, and Laura M. Sangalli

Non-intact Families and Children's Vaccination Coverage: Evidence from Italy 617
Maria Veronica Dorgali, Raffaele Guetto, Valentina Tocchioni, Alice Dominici, and Daniele Vignoli

Quantile Treatment Effect and GBM: An Approach Based on Propensity Score 623
Francesco Dotto, Francesco Giuli, Margherita Scarlato, and Francesco Bloise

High Volatility, High Emissions? A Hidden-Markov Model Approach 629
Pierdomenico Duttilo, Marina Bertolini, Alfred Kume, and Stefano Antonio Gattone

Mixtures of Generalized Latent Trait Analyzers for Biclustering Multivariate Data 635
Dalila Failli, Maria Francesca Marino, and Francesca Martella

Detecting Fake News from Text: A Stagewise Methodology 640
Matteo Farnè and Giulia Benelli

Examining Socioeconomic Inequalities in Healthy Working Life Expectancy and the Association with Smoking Behaviour 646
Alessandro Feraldi and Christian Dudel

Monitoring All-Cause and Cause-Specific Mortality Inequalities in Italian Regions 653
Alessandro Feraldi, Marilena Pappagallo, Cristina Giudici, and Luisa Frova

A Nonparametric Test for Fixed Effects in Crossed Random Effect Models 659
Federico Ferraccioli, Livio Finos, and Angela Andreella

A Bayesian Cure Model for Earthquake Parameter Estimation Using Crowdsourced Smartphone Data 665
Francesco Finazzi, Luca Aiello, and Lucia Paci

Analysis of Mastodon's Italian Messages: Networks, Topics, LLM and Machine Learning 671
Livio Finos and Andrea Sciandra

Author Index 677

Contributed Sessions I

Bayesian Inference for Generalized Linear Models via Quasi-Posteriors: an Application to Eurasian Chaffinch Abundance in Finland

Davide Agnoletto[1(✉)], Tommaso Rigon[2], and David B. Dunson[3]

[1] Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy
davide.agnoletto.2@phd.unipd.it
[2] Department of Economics, Management and Statistics, University of Milano–Bicocca, 20126 Milano, Italy
[3] Department of Statistical Science, Duke University, Durham, NC, USA

Abstract. Generalized linear models (GLMs) are commonly used to investigate relationships between a response variable and covariates, offering straightforward interpretability.However, concerns about model misspecification impacting inferential outcomes arise.An established frequentist technique, the quasi-likelihood, enhances robustness by necessitating specification of only the first two moments.We leverage on quasi-likelihoods in developing a robust approach for Bayesian inference in GLMs.Quasi-posteriors follow a coherent generalized Bayes approach and have desiderable large sample properties.In this paper we use the quasi-posterior for modeling the abundance of Eurasian chaffinch (*Fringilla coelebs*) in Finland in year 2014.

Keywords: Generalized Bayes · Model misspecification · Quasi-posterior · Robustness

1 Introduction

Generalized linear models ([1,2]) are a widespread statistical tool for modeling a large variety of data. Despite their popularity, assuming that the distribution of the response variable belongs to an exponential family could lead to model misspecification, compromising inferential conclusions. A common scenario arises when dealing with heteroscedastic continuous data, where the Gaussian linear model assumption of homoscedasticity is violated. Traditionally, the solutions adopt a model-based approach increasing the complexity of the model or, in alternative, a nonparametric approach that increases the flexibility avoiding to make too stringent assumptions. However, both these approaches can turn out to be problematic.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 3–7, 2025.
https://doi.org/10.1007/978-3-031-64431-3_1

In this paper we consider a third way adopting a semi-parametric approach that makes weaker distributional assumption on the response variable and simultaneously preserves interpretability and computational tractability. To do so we rely on the second-order conditions and inference on the parameter of interest is carried out through quasi-posterior [3], i.e. the update of prior beliefs using the quasi-likelihood function. The quasi-likelihood method [4], is a well-established technique within the frequentist literature and allows to specify only the mean and variance of the response variable instead of its entire distribution. In particular, let Y_i be independent random variables and $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^p$ be the observation-specific vector of covariates for $i = 1, \dots, n$, and let $\boldsymbol{\beta} \in \mathbb{R}^p$ be the parameter of interest.

Quasi-likelihood method assumes

$$\mathbb{E}(Y_i) = \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \quad \text{var}(Y_i) = \psi V(\mu_i), \tag{1}$$

where $g(\cdot)$ is the link function, $V(\cdot)$ is a generic variance function and $\psi > 0$ is the dispersion parameter.

For maintaining the formal validity of the inferential procedure, we embrace the generalized Bayesian paradigm, about which interest is recently incrasing (see [5]). Quasi-posteriors have relevant connections with the notion of coarsened posterior of [6] but leverage on the consolidated literature of quasi-likelihood to facilitate their practical implementation, especially for the choice of the coarsening parameter. Moreover, when $n \to +\infty$, the quasi posterior distribution converges in total variation to a Gaussian distribution and it is well calibrated in terms of frequentist coverage. Importantly, there exists a connection with the loss-likelihood bootstrap introduced by [7] that, together with the one-to-one relation with the dispersion parameter, allows to assign a value to the loss-scale parameter using a method of moments estimator.

2 Quasi-Posteriors

Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top \in \mathcal{Y}^n$ be a vector of random variables whose realization is $\boldsymbol{y}$, and $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_n^\top)$ be a matrix of covariates. Standard GLMs assume that observations y_i, $i = 1, \dots, n$, are realizations of $Y_i \mid \boldsymbol{x}_i$, whose distribution belongs to the exponential dispersion family. Here, instead, we rely on the second-order assumptions (1), which specify only the first two moments of $Y_i \mid \boldsymbol{x}_i$. This allows

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\psi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_r} = 0, \tag{2}$$

to coincide with the general likelihood equation for an exponential dispersion family. When the derivative of left-hand side of (2) with respect to $\boldsymbol{\beta}$ is positive definite, it is possible to specify the log-quasi-likelihood function $\ell_Q(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}, \psi) = \frac{1}{\psi} \sum_{i=1}^{n} \int_{y_i}^{\mu_i(\boldsymbol{\beta})} \frac{y_i - t}{V(t)} \mathrm{d}t$ having (2) as gradient but not necessarily corresponding to any distribution. The integral can be written in closed form

for several choices of variance function, including some known distributions as Gaussian, Poisson, and binomial. In this context [3] define the quasi-posterior as

$$p_Q(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \psi) \propto p(\boldsymbol{\beta}) \exp\left\{ \frac{1}{\psi} \sum_{i=1}^{n} \int_{y_i}^{\mu_i(\boldsymbol{\beta})} \frac{y_i - t}{V(t)} \mathrm{d}t \right\} \tag{3}$$

where $p(\boldsymbol{\beta})$ is the prior distribution for $\boldsymbol{\beta}$. The quantity $p_Q(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \psi)$ is a well-defined probability density function provided that the normalizing constant is positive and finite, and represents a proper update of a belief distribution, according to the notion of general Bayesian model of [5]. The target of quasi-posterior inference is $\boldsymbol{\beta}_{\text{OPT}} = \arg\max_{\boldsymbol{\beta}} \int \ell_Q(\boldsymbol{\beta}; y, \boldsymbol{x}, \psi) f_0(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y$, where $f_0(\boldsymbol{x}, y)$ denotes the true data generator, and coincides with the true parameter value $\boldsymbol{\beta}_0$ assumed by the quasi-likelihood approach.

Asymptotically, the quasi-posterior distribution of $\boldsymbol{\beta}$ is approximately a p-variate Gaussian with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix $I(\boldsymbol{\beta}_0)^{-1}$, where $\hat{\boldsymbol{\beta}}$ is the solution of (2) with respect to $\boldsymbol{\beta}$ and $I(\boldsymbol{\beta}_0) = \psi^{-1} \boldsymbol{X}^\top \tilde{\boldsymbol{D}} \boldsymbol{X}$, with $\tilde{\boldsymbol{D}} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$, $\tilde{d}_i = [V(\mu_i(\boldsymbol{\beta}_0)) g'(\mu_i(\boldsymbol{\beta}_0))^2]^{-1}$, for $i = 1, \ldots, n$. Thanks to this result, quasi-posteriors are well-calibrated in terms of frequentist coverage. This means that, asymptotically, the posterior credible sets of level ρ contain the true parameter value $\boldsymbol{\beta}_0$ with a probability of ρ.

Furthermore, asymptotic normality is a crucial outcome for assessing the scale of the quasi-posterior. The proportion of information derived from the quasi-likelihood and the prior is regulated by the loss-scale parameter ψ, which also controls the dispersion assumed by the second-order conditions within the quasi-likelihood approach. This connection offers new perspectives on the choice of this value, which is a delicate issue in the generalized Bayes literature [8]. Generalized Bayes procedures typically assign a value to the loss-scale parameter based on the data before inferring on $\boldsymbol{\beta}$. In our case, the quasi-posterior has important connections with the approach for assigning a value to ψ proposed by [7] based on the loss-likelihood bootstrap. Motivated by this connection and by the link between loss-scale parameter and dispersion, [3] propose to use the method of moments estimator

$$\hat{\psi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{4}$$

to quantify the value of ψ, where $\hat{\mu}_i = \mu_i(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ is the solution of (2) with respect to $\boldsymbol{\beta}$.

3 Application: *Fringilla Coelebs* abundance in Finland

We analyze data from an ecological monitoring study of birds, focusing on the occurrence rates of the Eurasian chaffinch (*Fringilla coelebs*) in Finland for the year 2014 [9]. The dataset consists of $n = 244$ observations collected from different sites. For each observation, the data include a set of environmental covariates

that provide information about the habitat type (classified as either broadleaved forest, coniferous forest, open habitat, urban habitat or wetlands) and the mean temperature registered in April-May. The chaffinch is a partially migrant passerine and is common in Finland during certain times of the year. Its abundance varies substantially across the country, and our focus is on assessing to what extent this variation is explained by temperature and habitat features.

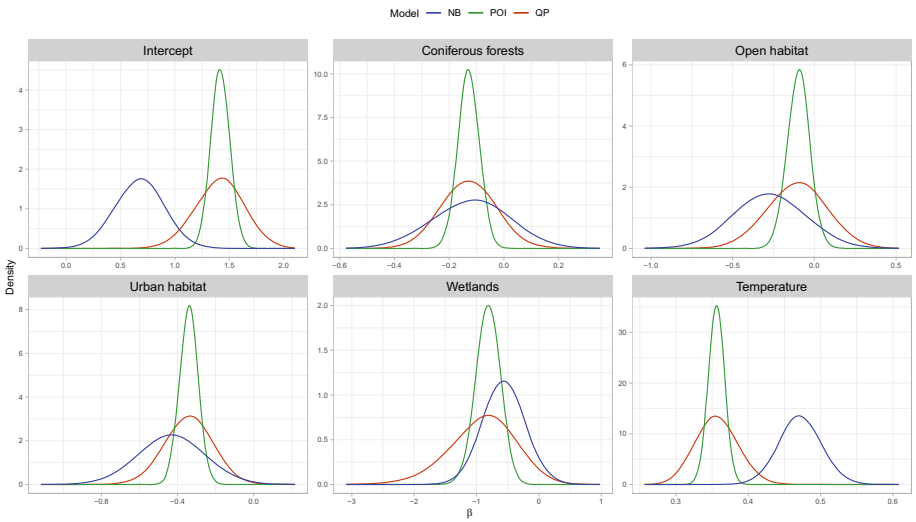


Fig. 1. Posterior distribution for regression coefficients.

Standard practice in modeling of bird species abundance data focuses on using Poisson GLMs. However, bird occurrence counts tend to be overdispersed; indeed, for the chaffinch data, the estimated dispersion is $\hat{\psi} = 6.95$, suggesting substantial overdispersion relative to the Poisson distribution. Here, we focus on the log link, as it leads to easy interpretation of regression coefficients and is standard in ecological abundance modeling. The covariate vector includes temperature, habitat, and year. Specifically, the temperature variable corresponds to the average temperature (in degrees Celsius) recorded during April and May at each site. The different habitats mentioned above are included as dummy indicators, with broadleaved forest considered as the baseline, as it is the primary native habitat of this species.

We compare our quasi-posterior with Poisson and negative binomial regression models. To assess goodness-of-fit, we compare the in-sample standardized mean squared error (sMSE) of the three models computed using Pearson residuals, which remove the effect of assumed heteroscedasticity. As expected, the Poisson model has the highest sMSE (6.61) since it underestimates the dispersion, while the quasi-posterior and negative binomial model perform substantially bet-

ter since both account for overdispersion, with a better fit for the quasi-posterior (0.95 sMSE vs 1.01 for the negative binomial).

Posterior means and 0.95 credible intervals for $\boldsymbol{\beta}$ are reported in Fig. 1. As expected, the negative binomial model and quasi-posterior provide similar estimates for the regression coefficients, and their posterior distributions have higher variance than those under the Poisson model. Eurasian chaffinch abundance is lower in urban habitats and wetlands compared to broadleaved forests, while there is not enough evidence to argue the same about coniferous forests. Moreover, *Fringilla coelebs* diffusion significantly varies according to the average spring temperature across the locations. This indicates that the temperature the birds encounter in Finland after their spring migration affects their spreading over the whole territory, which is influenced mainly by the habitat.

References

1. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. J. R. Stat. Soc. Series A (Gen.) **135**(3), 370–384 (1972)
2. McCullagh, P., Nelder J. A.: Generalized Linear Models. Chapman & Hall (1989)
3. Agnoletto, D., Rigon, T., Dunson, D.B.: Bayesian inference for generalized linear models via quasi-posteriors. arXiv:2311.00820
4. Wedderburn, R.W.: Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika **61**(3), 439–447 (1974)
5. Bissiri, P.G., Holmes, C.C., Walker, S.G.: A general framework for updating belief distributions. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **78**(5), 1103–1130 (2016)
6. Miller, J.W., Dunson, D.B.: Robust Bayesian inference via coarsening. J. Am. Stat. Assoc. **114**(527), 1113–1125 (2019)
7. Lyddon, S.P., Holmes, C., Walker, S.G.: General Bayesian updating and the loss-likelihood bootstrap. Biometrika **106** (2), 465–478 (2019)
8. Holmes, C.C., Walker, S.G.: Assigning a value to a power likelihood in a general Bayesian model. Biometrika **104**(2), 497–503 (2017)
9. Lindström, A., Green, M., Husby, M., Kalas, J.A., Lehikoinen, A.: Large-scale monitoring of waders on their boreal and arctic breeding grounds in northern Europe. Ardea **103**(1), 3–15 (2015)



A Spatial Product Partition Model for PM_{10} Data

Luca Aiello[1(✉)], Sirio Legramanti[2], and Lucia Paci[3]

[1] Department of Economics, Management and Statistics, Università degli Studi di Milano-Bicocca, Milano, Italy
l.aiello4@campus.unimib.it
[2] Department of Economics, Università degli studi di Bergamo, Bergamo, Italy
[3] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

Abstract. This work illustrates a model-based clustering method for analyzing PM_{10} measurements over time. In particular, we develop a Bayesian dynamic linear model coupled with a spatial product partition model for clustering monitoring stations that exhibit similar persistence and variability of the PM_{10} concentrations over time. The model integrates spatial information (the locations of the considered monitoring stations) into the clustering process in order to increase the probability that neighboring stations will be assigned to the same cluster.

This methodology is applied to the time series of daily PM_{10} measurements recorded by 110 monitoring stations in Austria. Our analysis reveals three spatially cohesive clusters characterized by different levels of persistence and variability of the PM_{10} concentrations. These results may provide helpful insights for understanding air pollution dynamics and support policymakers in identifying intervention areas.

Keywords: Bayesian nonparametrics · clustering · environmental statistics · spatiotemporal modeling

1 Introduction

Air quality is one of the most pressing threats to public health [11]. Particulate matter (PM) is a standard proxy indicator for air pollution. Rather than a single chemical, PM is a blend of solid and liquid particles that linger in the atmosphere for extended periods. These particles stem from various origins, including natural sources like soil erosion, volcanic activity, and pollen dispersion, alongside anthropic sources like industrial operations, heating systems, and vehicle emissions. For regulatory purposes, PM is classified by diameter, with a notable PM class being that of particles with a diameter of 10 microns or less (PM_{10}). Such fine particles are inhalable into human lungs, thus inducing adverse health effects.

PM_{10} levels are measured daily by monitoring stations, thus generating a consistent amount of spatiotemporal data. Such data for Europe are made publicly

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 8–13, 2025.
https://doi.org/10.1007/978-3-031-64431-3_2

available by the European Environmental Agency (https://www.eea.europa.eu/en). The European Commission's air quality standards limit PM_{10} concentrations to a daily average of 50 $\mu g/m^3$, but many European countries struggle to adhere to the allowed annual number of exceedances of the daily limit.

The goal of this work is to cluster monitoring stations based on the behavior of the time series associated with each station. To achieve this, we integrate a spatial Product Partition Model (sPPM) [8] into a Bayesian state-space model [10], customarily used to study PM_{10} dynamics. This approach not only enables the dissection of temporal trends in PM_{10} concentrations but also incorporates the spatial location of monitoring stations into our clustering model. This fusion allows us to identify clusters of stations whose time series share similar traits while facilitating that stations close to each other are clustered together. Such a spatial cohesion of clusters can be desirable from a policymaker's perspective, since it helps designing areas of intervention.

In this short paper, we focus on the Austrian data over 100 d in 2018 (see Sect. 3). Although limited in space and time, this case study allows us to showcase the potential of our Bayesian nonparametric technique (see Sect. 2).

2 Model

Let $y_t(\boldsymbol{s}_i)$ denote the PM_{10} level observed on day t at station i, which is located at $\boldsymbol{s}_i = (lon_i, lat_i)^\top$, for $t = 1, \ldots, T$ and $i = 1, \ldots, n$. Adopting a model similar to the one in [7], let

$$\begin{aligned} y_t(\boldsymbol{s}_i) &= \boldsymbol{z}_t^\top \boldsymbol{\alpha}_i + w_t(\boldsymbol{s}_i) + \epsilon_t(\boldsymbol{s}_i), \qquad \text{with} \quad \epsilon_t(\boldsymbol{s}_i) \overset{ind}{\sim} N\left(0, \sigma_i^2\right), \\ w_t(\boldsymbol{s}_i) &= \rho_i w_{t-1}(\boldsymbol{s}_i) + \nu_t(\boldsymbol{s}_i), \qquad \text{with} \quad \nu_t(\boldsymbol{s}_i) \overset{ind}{\sim} N\left(0, \tau_i^2\right), \end{aligned} \tag{1}$$

where $\boldsymbol{z}_t$ is a vector with four dummy variables, including an intercept, introducing a quadratic trend and taking into account the seasonality, $\boldsymbol{\alpha}_i$ is the associated vector of coefficients, and $w_t(\boldsymbol{s}_i)$ is a spatiotemporal random effect equipped with an autoregressive temporal structure.

To cluster the considered monitoring stations, we employ a sPPM as a prior for the parameters governing the autoregressive process $w_t(\boldsymbol{s}_i)$. Namely, we pose such a prior on the station-specific autocorrelation (ρ_i) and variability (τ_i^2), in order to cluster the stations based on their parameters, $\boldsymbol{\gamma}_i = (\rho_i, \tau_i^2)$.

Let $\phi_n = \{S_1, \ldots, S_{k_n}\}$ be a partition of the n stations into k_n subsets, and denote with $\boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ the complete sets of locations and PM_{10} data, respectively. Additionally, let $\boldsymbol{s}_h^\star = (\boldsymbol{s}_i, i \in S_h)$ and $\boldsymbol{y}_h^\star = (\boldsymbol{y}_i, i \in S_h)$ represent the locations and PM_{10} data of the stations in cluster h. Finally, denote by $g(\boldsymbol{s}_h^\star)$ a nonnegative function representing the similarity of the stations in cluster h with respect to their location $\boldsymbol{s}_h^\star$. Higher values of $g(\boldsymbol{s}_h^\star)$ correspond to sets of locations $\boldsymbol{s}_i$ closer to each other. Following [8], we define a sPPM by combining a PPM with cohesion $C(\cdot)$ with the similarity function $g(\cdot)$:

$$p\left(\phi_n \mid \boldsymbol{s}\right) \propto \prod_{h=1}^{k_n} g(\boldsymbol{s}_h^\star) C\left(S_h\right). \tag{2}$$

A popular choice for the cohesion is the one associated with the Dirichlet Process, i.e., $C(\cdot) = M \times \Gamma(|S_h|)$ with M being the scale parameter, Γ the gamma function, and $|S_h|$ the cardinality of cluster h. As for the similarity, [5] propose

$$g(\boldsymbol{s}_h^\star) = \int \prod\nolimits_{i \in S_h} q(\boldsymbol{s}_i \mid \boldsymbol{\xi}_h)\, q(\boldsymbol{\xi}_h)\, d\boldsymbol{\xi}_h. \tag{3}$$

In (3) there is no notion of the locations $\boldsymbol{s}_i$ being random variables, but the use of a probability density $q(\cdot)$ for the construction of $g(\cdot)$ facilitates computations. When the individual attributes $\boldsymbol{s}_i$ are continuous, as in our case, [5] recommend

$$q(\boldsymbol{s}_i \mid \boldsymbol{\xi}_h) = N(\boldsymbol{s}_i \mid \boldsymbol{m}_h, V_h), \qquad q(\boldsymbol{\xi}_h) = NIW(\boldsymbol{m}_h, V_h \mid \boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda_0}),$$

with $\boldsymbol{\xi}_h = (\boldsymbol{m}_h, V_h)$ being the mean and covariance matrix associated to cluster h, $N(\cdot,\cdot)$ the density of a normal distribution and $NIW(\cdot,\cdot,\cdot,\cdot)$ the density of a normal inverse-Wishart distribution.

We conclude the model specification by defining the prior distributions. For $i = 1, \ldots, n$, each $\boldsymbol{\alpha}_i$ is given an i.i.d. multivariate Gaussian prior with zero mean and diagonal covariance matrix $\boldsymbol{\Sigma_\alpha} = \text{diag}\left(\psi_1^2, \ldots, \psi_p^2\right)$, whose elements are assigned an inverse-Gamma prior. Concerning the pure error variances σ_i^2, for $i = 1, \ldots, n$, they are assigned independent inverse-Gamma priors. Moreover, the cluster-specific parameters $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_{k_n}^*$, such that $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_h^*$ for any station $i \in S_h$, are i.i.d. from $G_0 = \text{Beta}(a_\rho, b_\rho) \times \text{IG}(a_\tau, b_\tau)$. Finally, latent allocation variables $c_1, \ldots, c_n$ (such that $c_i = h$ if and only if $i \in S_h$) are given a multinomial prior with component probabilities derived from the sPPM [8].

We then sample from the posterior distribution of the model parameters via the Markov chain Monte Carlo (MCMC) Algorithm 8 in [6]. This algorithm employs l additional empty clusters, whose cluster-specific parameters (serving as auxiliary variables) are generated from the prior distribution. To be more specific, let $S_h^{(-i)}$ represent the h-th cluster, $k_n^{(-i)}$ denote the number of clusters, and $\boldsymbol{s}_h^{\star(-i)}$ be the location of the stations in cluster h, all when station i is not taken into account. Then, at each iteration of the algorithm, we assign station i to cluster h, conditional on all the other parameters, with probability

$$\Pr(i \in S_h \mid -) \propto \begin{cases} f(\boldsymbol{w}_i \mid \boldsymbol{\gamma}_h^*, -) \dfrac{C\left(S_h^{(-i)} \cup \{i\}\right) g\left(\boldsymbol{s}_h^{\star(-i)} \cup \boldsymbol{s}_i\right)}{C\left(S_h^{(-i)}\right) g\left(\boldsymbol{s}_h^{\star(-i)}\right)} & h = 1, \ldots, k_n^{(-i)} \\ f(\boldsymbol{w}_i \mid \boldsymbol{\gamma}_h^*, -) C(\{i\})\, g(\boldsymbol{s}_i) & h = k_n^{(-i)} + 1, \ldots, k_n^{(-i)} + l \end{cases} \tag{4}$$

where $\boldsymbol{w}_i = (w_1(s_i), \ldots, w_T(s_i))^\top$ and $f(\boldsymbol{w}_i \mid \boldsymbol{\gamma}_h^*, -) = N_T(\boldsymbol{0}, \boldsymbol{R}(\boldsymbol{\gamma}_h^*))$, with $[\boldsymbol{R}(\boldsymbol{\gamma}_h^*)]_{t,t'} = \tau_h^{*2} \rho_h^{*|t-t'|}$, is the probability density function of $\boldsymbol{w}_i$ evaluated at the parameters of cluster h.

3 Analysis of PM_{10} Data

We apply the model described in Sect. 2 to the daily PM_{10} concentrations measured from January 1, 2018 to April 10, 2018 by 110 monitoring stations

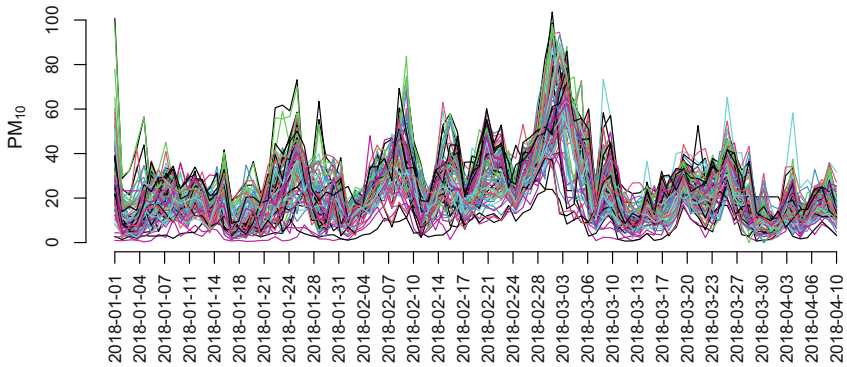


Fig. 1. Time series of daily PM_{10} concentrations recorded by 110 monitoring stations in Austria from January 1, 2018, to April 10, 2018.

in Austria. These data have been acquired from the European Environmental Agency and are publicly available at https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm.

Figure 1 illustrates the individual time series for each considered monitoring station within the specified time frame. High peaks are observed at the onset of the year and at the beginning of March across nearly all stations, with intermittent peaks throughout the time frame under consideration. Moreover, substantial variability can be observed in PM_{10} levels across the entire period.

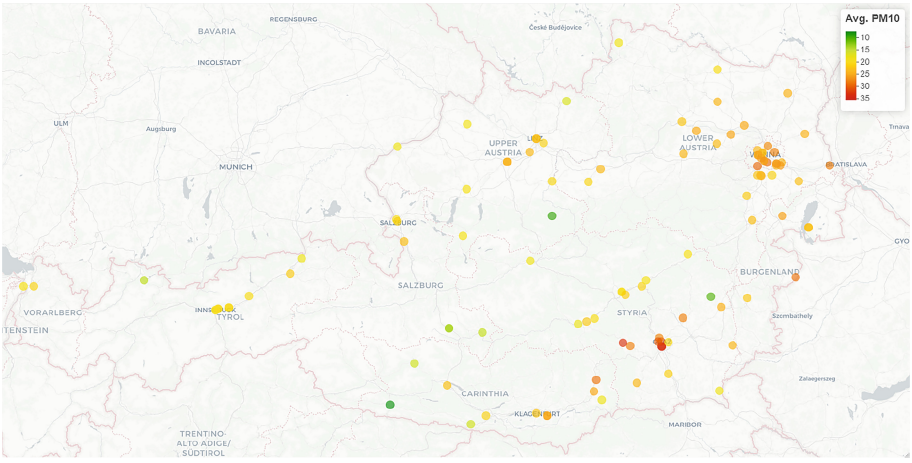


Fig. 2. Average PM_{10} concentrations measured at the considered monitoring stations.

Figure 2 depicts the considered stations on the map, color-coded based on the mean concentration of PM_{10} in the observational time frame. The highest PM_{10}

levels are observed around the four most populated Austrian cities: Vienna, Graz, Linz, and Salzburg. Specifically, Graz exhibits the highest levels, with Vienna following closely behind. Meanwhile, Linz and Salzburg display comparatively lower but still significant PM_{10} levels.

We ran the MCMC algorithm described in Sect. 2 for 5'000 iterations, discarding the first 1'000 ones as burn-in, and then retaining one every two of the remaining iterations. We thus obtained the posterior distribution of the cluster assignment of each station. We then employed a greedy, stochastic search approach [2] to obtain a posterior point estimate of the partition by minimizing a loss function. In particular, with the Binder loss function [1], we obtained three clusters, each associated with distinct $\gamma_h^* = (\rho_h^*, \tau_h^{2*})^\top$ values: $\boldsymbol{\gamma}_1^* = (0.33, 0.59)^\top$, $\boldsymbol{\gamma}_2^* = (0.47, 0.72)^\top$, and $\boldsymbol{\gamma}_3^* = (0.61, 0.41)^\top$; these values correspond to the average cluster-specific parameters at the last MCMC iteration.

The resulting three clusters are displayed on the map of Austria in Fig. 3. The first cluster contains stations with the lowest autocorrelation, the second cluster encompasses stations with the highest variability, and the third cluster is characterized by the strongest autocorrelation and the smallest variability. In other words, the third cluster (in red in Fig. 3) consists of monitoring stations where PM_{10} levels remain quite stable over time, with minimal fluctuations. For example, it is known that the PM_{10} persistence tends to be higher near urban areas and arid regions, while it tends to be lower in areas with more vegetation and increased air circulation [4,12]. Finally, note that, although overlapping, the obtained clusters exhibit some spatial cohesion, facilitating their interpretation and potential usage as areas of intervention. This cohesion is facilitated by the spatial similarity we introduced in the sPPM in Sect. 2.

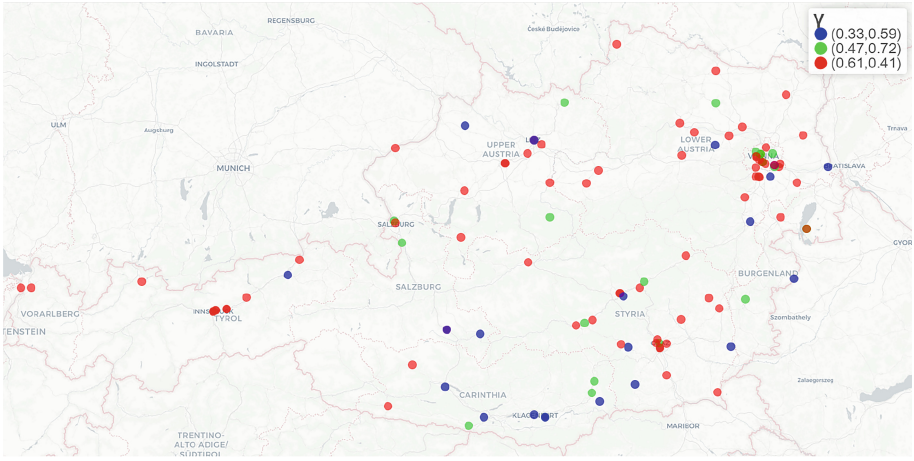


Fig. 3. Austrian monitoring stations colored according to their estimated cluster membership; cluster-specific parameters (ρ_h^*, τ_h^{2*}) are reported in the legend.

4 Conclusions

In this paper, we proposed a Bayesian clustering algorithm for analyzing environmental data and applied it to the time series of daily PM_{10} concentrations at 110 monitoring stations in Austria. Namely, we devised a hierarchical linear mixture model enriched with a first-order autoregressive process to capture temporal dynamics. By incorporating a Dirichlet Process prior for the autoregressive model parameters, we were able to cluster together monitoring stations with similar persistence and variability of PM_{10} over time. We also leveraged spatial information to facilitate that neighboring stations are clustered together. Our analysis revealed higher persistence in urban areas, contrasting with diminished persistence in the countryside and in the mountains.

Our model could be enriched by including spatiotemporal covariates in the mean structure [3]. For instance, weather information on a daily scale, such as temperature, precipitation, and wind speed, can be added to the observational equation in (1). Moreover, a possible future research direction concerns how the partition of the monitoring stations may change over time [9]. This could offer valuable insights in terms of air quality assessment.

References

1. Binder, D.A.: Bayesian cluster analysis. Biometrika **65**(1), 31–38 (1978)
2. Dahl, D.B., Johnson, D.J., Müller, P.: Search algorithms and loss functions for Bayesian clustering. J. Comput. Graph. Stat. **31**, 1189–1201 (2022)
3. Lagmiri, S., Dahech, S.: Weather Types and Their Influence on PM $_{10}$ and O_3 Urban Concentrations in the Cergy-Pontoise Conurbation. J. Appl. Meteor. Climatol. **62**, 549–561 (2023)
4. Meraz, M., Rodriguez, E., Femat, R., Echeverria, J.C., Alvarez-Ramirez, J.: Statistical persistence of air pollutants (O_3, SO_2, NO_2 and PM_{10}) in Mexico City. Physica A: Stat. Mech. Appl. **427**, 202–217 (2015)
5. Müller, P., Quintana, F., Rosner, G.L.: A product partition model with regression on covariates. J. Comput. Graph. Stat. **20**(1), 260–278 (2011)
6. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)
7. Nieto-Barajas, L.E., Contreras-Cristán, A.: A Bayesian nonparametric approach for time series clustering. Bayesian Anal. **9**(1), 147–170 (2014)
8. Page, G.L., Quintana, F.A.: Spatial product partition models. Bayesian Anal. **11**(1), 265–298 (2016)
9. Paci, L., Finazzi, F.: Dynamic model-based clustering for spatio-temporal data. Stat. Comput. **28**, 359–374 (2018)
10. West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models, 2nd edn. Springer, New York (1999). https://doi.org/10.1007/b98971
11. World Health Organization, Ambient (outdoor) air pollution, (2022). http://www.who.int/mediacentre/factsheets/fs313/en/
12. Zafra, C., Ángel, Y., Torres, E.: ARIMA analysis of the effect of land surface coverage on PM_{10} concentrations in a high-altitude megacity. Atmos. Pollut. Res. **8**(4), 660–668 (2017)



Efficient Bayesian Estimation of Spatial Poisson Auto-regression with Leroux Random Effects

Pierfrancesco Alaimo Di Loro[1(✉)], Dankmar Böhning[2], and Sujit K. Sahu[2]

[1] LUMSA University, Rome, Italy
p.alaimodiloro@lumsa.it
[2] University of Southampton, Southampton, UK

Abstract. The standard Poisson Auto-Regression framework considers static coefficients and does not incorporate any spatio-temporal dependence on the parameters governing the process dynamic. However, unobserved space-time variability is a very relevant component when dealing with observations organised in space and time. We consider a more flexible specification that can adjust for local deviations from the general pattern while borrowing information from adjacent areas. The model, specified in a Bayesian framework, might suffer from computational bottlenecks that can make its estimation unfeasible. Therefore, we implement it in STAN to jointly update all the parameters and improve mixing, while adopting a novel sparse-matrix representation to attain improved computational performances. The computational advantage and the model performances have been validated through a simulation study.

Keywords: Poisson autoregression · Spatio-temporal modelling · Leroux

1 Introduction

Poisson auto-regression is a very general framework for modelling time-series of count data that exhibit substantial temporal dependence patterns [1,2]. At the top level, the model assumes that each count is a realisation of a Poisson random variable whose rate is decomposed into the sum of an auto-regressive term (epidemic component) and a baseline (endemic component). Typical applications of this model assume a stationary process with a *static* behaviour, where the parameters are not allowed to vary across units or over time. More generally, we are not aware of any attempt to account for a spatio-temporally structured variability in the parameters. To fill this gap in the literature, [3] embeds the Poisson Auto-Regression into a Bayesian spatio-temporal framework that induces cross-dependence in the auto-regressive and baseline parameters by perturbing them with two sets of random effects endowed with a spatio-temporal Leroux prior [4]. Therefore, each observation is characterized by the specific regime of the process in that location at that time. Computational difficulties in the Bayesian

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 14–19, 2025.
https://doi.org/10.1007/978-3-031-64431-3_3

estimation MCMC routine may arise when the data size grows large in the number of time-points T and, most especially, of spatial locations L. This is further exacerbated by the presence of two high-dimensional sets of random effects that must be identified separately and can trouble the mixing of the chains. Therefore, attaining improved computational performance for evaluating the posterior under the Leroux prior is pivotal to making the model estimation scalable. This can be achieved by exploiting the sparsity of the Leroux prior and considering efficient strategies for the determinant evaluation. Furthermore, we propose estimating the model through the joint updates of all the parameters guaranteed by the most recent Hamiltonian Monte-Carlo dynamics [5], which are easily accessible through the Bayesian computing software STAN [6].

2 The Model

Let $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_T]$ be the $L \times T$ matrix of counts observed at time-points $t = 1, \dots, T$, with each column $\boldsymbol{y}_t = [y_{1t}, \dots, y_{Lt}]$ being the spatial vector spanning the L regions in $\mathcal{S} = \{\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_L\}$. These can be counts of events occurring in each region ℓ at each time t and, under Poisson auto-regression, they are assumed to be the realisations of a Poisson random variable $Y_{\ell t} \,|\, \boldsymbol{y}_{\ell(1:t-1)} \sim Poi(\lambda_{\ell t})$ with rate decomposed into the sum of an auto-regressive and a baseline term $\lambda_{\ell t} = r_\ell \cdot y_{\ell(t-1)} + b_\ell$, where $r_\ell \geq 0$ and $b_\ell \geq 0$ are location-specific parameters. The first regulates the impact of the previous count on the current one and it determines whether the process is stationary ($r_\ell < 1$) or not ($r_\ell \geq 1$). The latter is a baseline that governs the rate of spontaneous generations and ensures that the process will not die out with probability 1.

This model specification can describe a plethora of correlated counts time-series that display a mixture of endemic and epidemic behaviour [7–10].

We consider the possibility that these two behaviours vary stochastically in space and time by introducing two sets of random effects. The rate at location $\ell \in \mathcal{S}$ and time $t = 1, \dots, T$ is specified as:

$$\lambda_{\ell t} = y_{\ell(t-1)} \cdot r_{\ell t} + b_{\ell t}, \tag{1}$$

where:

$$\log(\widetilde{r}_{\ell t}) = \boldsymbol{x}_{\ell t}^\top \cdot \boldsymbol{\beta} + \phi_{\ell t}, \quad \log(b_{\ell t}) = \log(\text{off}_\ell) + \boldsymbol{v}_{\ell t}^\top \cdot \boldsymbol{\eta} + \psi_{\ell t}. \tag{2}$$

$\boldsymbol{x}_{\ell t}$ and $\boldsymbol{v}_{\ell t}$ are $(k+1) \times 1$ and $(\nu+1) \times 1$ vectors of covariates and intercept, $\boldsymbol{\beta} = [\beta_0, \dots, \beta_k]$ and $\boldsymbol{\eta} = [\eta_0, \dots, \eta_\nu]$ are the corresponding vectors of coefficients, $\{\phi_{\ell t}\}_{\ell t}^{LT}$ and $\psi_{\ell t}$ are two independent sets of space-time random effects, off_ℓ is an area-specific offset. We here consider the space-time extension of the Leroux model [11] originally proposed by [4] that we denote as *CAR-AR Leroux*. Let $\phi_{\ell t}$, $\ell \in \mathcal{S}$, $t = 1, \dots, T$ be a set of *Gaussian* space-time random effects over the discrete domain $\mathcal{S} = \{\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_L\}$ introduced in Eq. (2). Let $\boldsymbol{W}$ be a $L \times L$ adjacency matrix for the locations in $\mathcal{S}$, whose elements w_{ij} are equal to 1 if

and only if $i \neq j$ and $i \sim j$ (location $\boldsymbol{\ell}_i$ is a neighbour of location $\boldsymbol{\ell}_j$), and 0 otherwise. The number of neighbours of each location $\boldsymbol{\ell}_i$ is $N_i = \sum_{j=1}^{l} w_{ij}$.

The model proposed by [11] assumes $\boldsymbol{\phi}_t \sim \mathcal{N}_L\left(\mathbf{0},\ \sigma^2 \cdot \boldsymbol{Q}(\alpha,\ \boldsymbol{W})^{-1}\right)$ where

$$\boldsymbol{Q}(\alpha,\ \boldsymbol{W}) = (\alpha \cdot (\boldsymbol{D} - \boldsymbol{W}) + (1-\alpha)\boldsymbol{I}_L)$$

and $\sigma^2 > 0$. The space-time extension by [4,12] connects the T time slices through a first-order auto-regressive structure:

$$\begin{aligned} &\boldsymbol{\phi}_1 \sim \mathcal{N}_L\left(\mathbf{0},\ \sigma^2 \cdot \boldsymbol{Q}\left(\alpha, \boldsymbol{W}\right)^{-1}\right) \\ &\boldsymbol{\phi}_t \,|\, \boldsymbol{\phi}_{t-1}, \ldots, \boldsymbol{\phi}_1 \sim \mathcal{N}_L\left(\rho \cdot \boldsymbol{\phi}_{t-1},\ \sigma^2 \cdot \boldsymbol{Q}\left(\alpha, \boldsymbol{W}\right)^{-1}\right),\ t = 2, \ldots, T, \end{aligned} \tag{3}$$

where $0 < \rho < 1$ is the temporal auto-regressive coefficient and regulates the amount of temporal dependence.

2.1 Efficient Implementation

The implementation of the Leroux model for large networks ($L > 100$), in the space-time setting especially, can be computationally intensive. Indeed, the evaluation of the corresponding density encompasses the computation of a quadratic form on the $L \times L$ matrix $Q(\alpha, \boldsymbol{W})$ and its determinant, whose naive implementation has a cost scaling with $\mathcal{O}(L^2)$. A full MCMC run can become rapidly unfeasible for modest spatial sizes already. This bottleneck, typical of many other CAR specifications [13,14], requires the adoption of ad-hoc strategies to estimate the model in a reasonable amount of time. For the quadratic form, we can exploit that only few entries of $\boldsymbol{W}$, and hence of $\boldsymbol{Q}(\alpha, \boldsymbol{W})$, are non-zero. Adopting a sparse representation of $\boldsymbol{W}$ reduces the operations to only the non-zero entries and can significantly speed up the evaluation of the quadratic form. Additionally, borrowing from [15], we introduce a strategy to efficiently evaluate the determinant of $\boldsymbol{Q}(\alpha, \boldsymbol{W})$. Indeed:

$$Q(\alpha, \boldsymbol{W}) \ = \ \left(\boldsymbol{I}_L - \alpha\left(\boldsymbol{W} + \boldsymbol{I}_L - \boldsymbol{D}\right)\right). \tag{4}$$

Let $\lambda_i = 1, \ldots, L$ be the eigenvalues of $(\boldsymbol{W} + \boldsymbol{I}_L - \boldsymbol{D})$. The determinant of $Q(\alpha, \boldsymbol{W})$ can be obtained as:

$$|Q(\alpha, \boldsymbol{W})| \ \propto \ \prod_{i=1}^{L}(1 - \alpha\lambda_i).$$

The eigenvalues λ_i do not vary from iteration to iteration and can be pre-computed, requiring only the evaluation of the product of L terms (due to the update of α) to be executed at each iteration ($\mathcal{O}(L^2) \rightarrow \mathcal{O}(L)$).

We implemented this strategy in the STAN software [6] through a non-centered parametrization of the space-time random effects:

$$\begin{aligned} &\boldsymbol{\phi}_t^* \ \sim \mathcal{N}\left(\mathbf{0}\,,\, \boldsymbol{Q}\left(\alpha, \boldsymbol{W}\right)^{-1}\right), \quad t = 1, \ldots, T \\ &\boldsymbol{\phi}_1 \ = \ \boldsymbol{\phi}_1^*, \quad \boldsymbol{\phi}_t \ = \ \rho \cdot \boldsymbol{\phi}_{t-1} + \sigma \cdot \boldsymbol{\phi}_t^*, \quad t = 2, \ldots, T. \end{aligned}$$

This facilitates the MCMC sampling of the random effects in the *unconstrained* parameters space used by STAN, reducing unnecessary correlation while discharging some additional (negligible) burden on the (now efficient) posterior computation [5]

Code and examples for the sparse and efficient implementation of the Leroux AR, together with examples of implementation of our Poisson AR model, are publicly available on GitHub at https://github.com/PAlaimo/PoiAR_SparseLeroux.

3 Simulation Study

We consider the $L = 313$ *Local Authority Districts, Counties and Unitary Authorities* (LADCUA) as a spatial reference and simulate data from a purely spatial model with Leroux random effects to compare the execution time of the naive Leroux implementation with our sparse version. We consider different (connected) subsets of all the areas with increasing sizes and evaluate the computing time of 1000 iterations for both approaches. Figure 1 highlights the different rates at which the cost of the procedure scales, with the sparse implementation yielding a sensible reduction in the run time. In particular, when $L = 313$ and $T = 1$ the naive Leroux requires ≈ 1 *hour* and our implementation less than 1 *minute* to complete $M = 1000$ iterations.

We also perform a full simulation study to test the proposed model on the 33 districts belonging to the region of *London* (see Fig. 2). $B = 400$ different sets of data for $T = 10$ time-points have been simulated. Each simulated set has two covariates $X_1, X_2 \sim \mathcal{N}(0, 1)$ on the auto-regressive coefficient and one covariate $V_1 \sim \mathcal{N}(0, 1)$ on the baseline. The parameters are also simulated independently in each set, with $\beta_0^* \sim \mathcal{N}(-1, 0.5)$, $\beta_j^* \sim \mathcal{N}(0, 0.5)$, $j \neq 0$, $\eta_j^* \sim \beta_j^* \sim \mathcal{N}(0, 1)$, $\alpha_\phi, \alpha_\psi \sim Unif(0, 1)$, $\rho_\phi, \rho_\psi \sim Unif(0, 1)$, $\sigma_\phi \sim \mathcal{N}^+(0, 0.1), \sigma_\psi \sim \mathcal{N}^+(0, 0.5)$. We keep 80% randomly chosen observations in-sample and use the remaining 20% for testing. We first evaluate the overall performances in terms of the true parameters recovery. Table 1 reports the *root Mean Squared Error* ($\overline{\text{rMSE}}$) and the *average coverage* ($\overline{\text{Cov}}_{95}$) of the corresponding 95% credibile intervals. The average rMSE is low in comparison to the average magnitude of each parameter, with coverages very close to their nominal levels. The performances in the

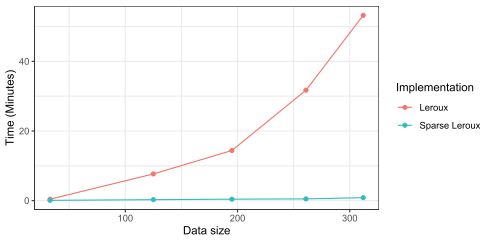


Fig. 1. Run time comparison of 1000 STAN by Leroux implementation

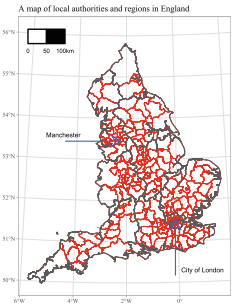


Fig. 2. Map of English LADCUAs

Table 1. Root mean square error and average coverage of the 95% credible intervals of all the parameters over the B simulated sets

Metric	β_0	β_1	β_2	η_0	η_1	α_ϕ	α_ψ	ρ_ϕ	ρ_ψ	σ_ϕ	σ_ψ
$\bar{\text{rMSE}}$	0.04	0.02	0.03	0.08	0.10	0.19	0.14	0.18	0.12	0.04	0.05
$\bar{\text{Cov}}_{95}$	0.94	0.96	0.95	0.93	0.94	0.94	0.95	0.98	0.93	0.94	0.95

retrieval of the latent space-time random effects are evaluated globally according to the same two metrics, with $\bar{\text{rMSE}} = 0.07$ and $\bar{\text{Cov}}_{95} = 0.93$ for $\phi_{\ell t}$ and $\bar{\text{rMSE}} = 0.23$ and $\bar{\text{Cov}}_{95} = 0.95$ for $\psi_{\ell t}$.

The prediction performances have been verified both *in* and *out-of* sample. The in-sample performances yield $\bar{\text{RMSE}} = 0.02$ and $\bar{\text{Cov}}_{95} = 0.99$, while the out-of-sample performances $\bar{\text{RMSE}} = 0.08$ and $\bar{\text{Cov}}_{95} = 0.96$ out-of-sample. The RMSE is clearly larger *out-of-sample* than *in-sample*, but the values are both low and, most importantly, the coverage is always close to the nominal level.

4 Conclusions

This work describes the space-time extension of the Poisson Auto-regressive model by [3] and draws attention onto the the computational strategies to make the model estimation feasible. The objective is achieved by re-writing the Leroux prior precision matrix and exploiting linear algebra to reduce the burden of the posterior computation. Furthermore, the non-centered parametrization of the CAR-AR Leroux in STAN allows the efficient joint sampling of the parameters in their high-dimensional space while guaranteeing the full exploration of the posterior. The improved computational performances are verified on a purely spatial example, where the different computational costs are discussed and highlighted. Furthermore, the recovery of all parameters, latent processes included, is satisfactory and shows how the model is indeed able to disentangle the auto-regressive from the baseline components. Furthermore, the predictive performances are more than satisfactory. This work can be extended in multiple

directions. For instance, a more general specification would envision the possibility of a correlation between the two (now independent) sets of random effects. This is easily done on paper, but it is key to find alternative strategies to keep the procedure efficient (i.e. evaluate efficiently the inverse and determinant). On another note, deciding what covariates to include in the auto-regressive and the baseline linear predictors is demanded of the researcher's sensibility. It would be extremely beneficial to introduce an automatic variable selection scheme that could facilitate such a decision process.

References

1. Kedem, B., Fokianos, K.: Regression Models for Time Series Analysis. John Wiley & Sons, New York (2005)
2. Fokianos, K., Rahbek, A., Tjøstheim, D.: Poisson autoregression. J. Am. Stat. Assoc. **104**(488), 1430–1439 (2009)
3. Alaimo Di Loro, P., Boehning, D., Sahu,S.: A Bayesian spatio-temporal extension to poisson auto-regression: modeling the disease infection rate of COVID-19 in England (2023)
4. Rushworth, A., Lee, D., Mitchell, R.: A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. Spatial spatio-temporal epidemiology **10**, 29–38 (2014)
5. Betancourt, M., Girolami, M.: Hamiltonian monte carlo for hierarchical models. Curr. trends Bayesian method. appl. **79**(30), 2–4 (2015)
6. Stan Development Team: Stan modeling language users guide and reference manual **2**, 26 (2021)
7. Matteson, D.S., McLean, M.W., Woodard, D.B., Henderson, S.G.: Forecasting emergency medical service call arrival rates. Ann. Appl. Stat. **5**(2B), 1379–1406 (2011)
8. Zhu, F., Liu, S., Shi, L.: Local influence analysis for Poisson autoregression with an application to stock transaction data. Stat. Neerl. **70**(1), 4–25 (2016)
9. Agosto, A., Cavaliere, G., Kristensen, D., Rahbek, A.: Modeling corporate defaults: poisson autoregressions with exogenous covariates (PARX). J. Empir. Financ. **38**, 640–663 (2016)
10. Xu, X., Chen, Y., Chen, C.W., Lin, X.: Adaptive log-linear zero-inflated generalized Poisson autoregressive model with applications to crime counts. Annals Appl. Stat. **14**(3), 1493 – 1515 (2020)
11. Leroux, B.G., Lei, X., Breslow, N.: Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials, pp. 179–191. Springer (2000). https://doi.org/10.1007/978-1-4612-1284-3
12. Rushworth, A., Lee, D., Sarran, C.: An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **66**(1), 141–157 (2017)
13. Cressie, N.: Statistics for Spatial Data. John Wiley & Sons, New York (2015)
14. Mingione, M., et al.: Spatio-temporal modelling of COVID-19 incident cases using Richards' curve: an application to the italian regions. Spatial Stat. **49**, 100544 (2022)
15. Jin, X., Carlin, B.P., Banerjee, S.: Generalized hierarchical multivariate CAR models for areal data. Biometrics **61**(4), 950–961 (2005)



A Two-Step Approach to Analyze the Inclusivity Distribution in Childcare Services in Italy

Angela Andreella[2], Gaia Bertarelli[1(✉)], Federico Caldura[1], and Stefano Campostrini[1]

[1] Department of Economics, Ca' Foscari University of Venice, Venice, Italy
gaia.bertarelli@unive.it
[2] Department of Economics and Management, University of Trento, Trento, Italy

Abstract. This paper studies inclusiveness in Early Childhood Education and Care (ECEC) services in Italy. The Italian legislative decree n. 65/2017 emphasizes the inclusiveness of ECEC services, defined as openness to all children. We propose a two-step model to analyze this concept. In the first step, the inclusiveness is estimated as a latent variable through an item response theory model. Then, the empirical best prediction (EBP) approach provides a reliable aggregate estimate of the phenomenon at the regional (NUTS2) level, which are unplanned domains for the considered data. In the second step, a mixed quantile model analyzes the distribution of this new inclusivity index, showing the disparities between public and private ECEC services and identifying virtuous regional scenarios. The results underline the need to standardize the accessibility features of ECEC services across Italy.

Keywords: Inclusiveness · ECEC · IRT model · mixed quantile regression

1 Introduction

Early Childhood Education and Care (ECEC) services are crucial in cultivating well-rounded economic and social conditions within countries [1], promoting gender employment equality [2], addressing demographic challenges [3] and improving the cognitive development of children [4]. In Italy, these services were introduced in 1971 and expanded in the 1990 s. Italy generally still lacks ECEC services, with an overall coverage falling below the European threshold set at 33% and distant from the new target set by EU for the 2023 of 45%. This lack is unequal across Italy, highlighting regional disparities. Addressing these disparities is crucial for improving ECEC service coverage in Italy [5,6]. Moreover, a pivotal aspect for policymakers is understanding how much of the limited and variable supply is truly accessible to the population. Even with a considerable number of available places, if they are not easily accessible, the overall situation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 20–26, 2025.
https://doi.org/10.1007/978-3-031-64431-3_4

for the country will not improve. The Italian legislative decree n. 65/2017 defines ECEC services as inclusive if open to all children, respecting the individuality, culture, and religion of the child and their family.

The Italian National Statistical Institute (Istat) [7,8] has analyzed preliminary data on this inclusivity concept, underscoring a notable territorial fragmentation. Public ECEC services are uneven across the country [7], more prevalent in the north and large cities. This is important since most ECEC services providing financial aid for families facing socio-economic circumstances (e.g., low-income, foreign, children with disabilities) are public [8]. Italy is characterized by various experiences and local choices affecting management models, fee reduction mechanisms, ranking criteria, and accommodation of children with disabilities. The fragmented landscape raises concerns about neglecting inclusiveness as a harmonized and systemic attribute. It is crucial to expand ECEC services, ensuring not only an increase in terms of available places but also in terms of inclusivity and accessibility for families, regardless of their geographical, economic, and social situations. Analyzing the concept of inclusivity requires novel data and proper statistical approaches that can synthesize its multidimensionality and highlight the ECEC services features.

This work proposes a two-step model to define and analyze the concept of inclusivity. Based on the available data, we assume that two dimensions can define an ECEC service as inclusive. The first describes the social aspect, while the second is the economic aspect. Then, in the first step, we estimate an index of inclusivity as a latent trait using an Item Response Theory model (IRT) [9] and return its estimates at the regional level using the empirical best prediction (EBP) approach [10] since regions are unplanned domains of the considered survey. To consider the within-region variability and the bimodal distribution of the inclusivity index, in the second step, we estimate a mixed quantile model [11]. Thanks to this two-step model, we found some virtuous regions where private childcare services are also characterized by a high level of inclusivity.

2 Data

The proposed analysis is based on the 2022 Italian sample survey on ECEC services [8], which recollects information such as the families' demand, accommodation capacity, occupancy rates, and quality of the educational offer for children interviewing ECEC services. This paper focuses on the survey section devoted to measuring the accessibility in terms of inclusiveness of the educational offer. Universally defining this concept may lead to discussions and debates on social, philosophical, and economic ideas, which is beyond this paper's scope. We practically define the concept of inclusiveness by two dimensions: a social and an economic one, starting from the available data and arguing that the selected variables described below can represent only one aspect of the multidimensional concept of inclusiveness.

Table 1 describes the variables used in our analysis. As mentioned earlier, we define inclusiveness by exploring the social (i.e., first 5 variables) and economic

Table 1. Description of the variables analyzed coming from [8].

Variable	Description
Foreign 0.25	1 if the $F_p \geq \mathcal{Q}_F(0.25)$, 0 otherwise
Foreign 0.5	1 if the $F_p \geq \mathcal{Q}_F(0.5)$, 0 otherwise
Foreign 0.75	1 if the $F_p \geq \mathcal{Q}_F(0.75)$, 0 otherwise
Foreign 1	1 if the $F_p \geq \mathcal{Q}_F(1)$, 0 otherwise
Disability	1 If at least one disabled person registered, 0 otherwise
Meal	1 If meal is included in the fee, 0 otherwise
Fee	1 If no entry fee, 0 otherwise
Disability fee	1 a reduction for child with a disability exists, 0 otherwise
Full Fee	1 If full tuition exemption is granted, 0 otherwise
ISEE	1 If there is a reduction for ISEE, 0 otherwise
Child	1 If there is a reduction for other children (enrolled or not), 0 otherwise
Social Services	1 If there is a reduction for social services indications, 0 otherwise
Family	1 If there is a reduction for another family condition, 0 otherwise

(i.e., last 7 variables) aspects of Italian ECEC services. As seen from Table 1, the questionnaire responses are thus simplified as dummy variables, losing some of the information but being able to handle the strong data variability. Moreover, the "foreigners" variable was constructed to consider the actual presence of foreigners in each territory. Let define by c_k the number of foreign children enrolled in the kindergarten k, where $k = 1, \ldots, 1323$ having 1323 childcare services surveyed, and by f_p where $p \in \{1, \ldots, 110\}$ the number of resident foreigners aged 0 to 2 years at the province level (i.e., sub-regional territory). The provincial foreign child enrolled rate is then defined as $F_p = \sum_{k \in \mathcal{P}_p} c_k / f_p$, where $\mathcal{P}_p$ is the set containing the childcare services index in the province p.

The foreign variable at level λ for the kindgarten $k \in \mathcal{P}_p$ is then defined as:

$$\text{Foreign } \lambda = \begin{cases} 1 & \text{if } F_p \geq \mathcal{Q}_F(\lambda) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{Q}_F(\lambda)$ is the quantile of $F = \{f_1, \ldots, f_{110}\}$ at level $\lambda \in \{0.25, 0.5, 0.75\}$. Therefore, the variable Foreign λ is equal for every $k \in \mathcal{P}_p$.

3 Inclusivity Index

In the first step of our analysis, an IRT model is employed to estimate the concept of inclusivity using the binary variables defined in Table 1.

Consider the manifest binary variable X_i where $i = 1, \ldots, 12$ following a Bernoulli distribution with expected value $\pi_i(\mathbf{z})$. Since this paper is interested in defining an indicator of inclusiveness, we assume that the concept of inclusivity is described by only one latent variable $\mathbf{z}$. The IRT model is defined by:

$$\log(\pi_i(\mathbf{z})/(1 - \pi_i(\mathbf{z}))) = \beta_{0i} + \beta_{1i}\mathbf{z} \tag{1}$$

where β_{0i} can be interpreted as the prevalence of the manifest variable i, and β_{1i} as the effect of the manifest variable i in the inclusivity definition.

The sampling weights at ECEC service level are used inside the likelihood-based estimation process of the parameters β_{0i}, β_{1i} and $\mathbf{z}$ [12].

Figure 1 shows the estimated probabilities $\hat{\pi}(\mathbf{z})$ across different values of the estimated latent variable, i.e., the inclusivity index. The probabilities $\hat{\pi}_i(\mathbf{z})$ are estimated for each variable i, i.e., the binary variables described in Table 1.

The most prevalent inclusive feature is the the meal included in the fee and the less prevalent one is the reduction for disabled children. The most impactful characteristic in defining an ECEC service inclusive is the ISEE reduction, while the least important is the meal's presence within the tuition.

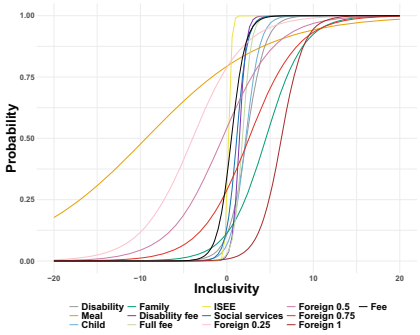


Fig. 1. Estimated probabilities $\hat{\pi}(\mathbf{z})$ across $\mathbf{z}$ for each covariate i defined in Table 1.

We scale $\hat{z}_k$ for each ECEC service k: $\tilde{z}_k = \hat{z}_k - \min_k \hat{z}_k / \max_k \hat{z}_k - \min_k \hat{z}_k$.

Two aspects of $\tilde{\mathbf{z}}$ are pointed out by the two figures below: (i) bimodality, (ii) high within-region variability. Figure 2 shows the frequency distribution of $\tilde{\mathbf{z}}$ by titularity. Figure 3 displays the median and interquartile range (IQR) values of $\tilde{\mathbf{z}}$ at the province level. These aspects lead us to study the features that define the bimodal distribution of the phenomenon with mixed quantile regression. Also, the high variability requires using a small area estimation approach to provide estimates of our inclusiveness index, at least at the regional level.

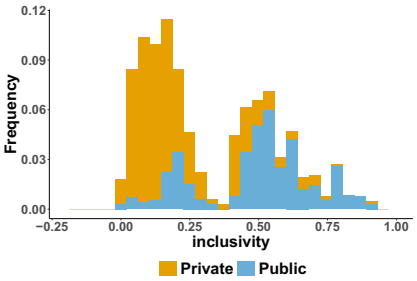


Fig. 2. Relative frequency distribution of the scaled inclusivity index estimated by the IRT model defined in Eq. (1) divided by type of service (i.e., private and public ones).

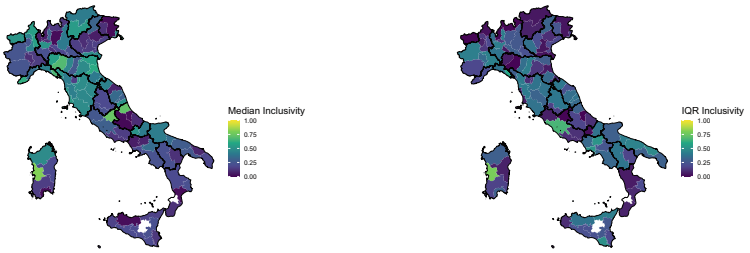


Fig. 3. The median and IQR of $\tilde{\mathbf{z}}$ estimated by the IRT model defined in Eq. (1) for each Italian province where the black lines represent the regional boundaries.

4 Virtuos Scenarios

Let $\tilde{\mathbf{z}} = [\tilde{z}_{ij}] \in \mathbb{R}^k$, where $i = 1, \ldots, n_j$ is the single ECEC service, and $j = 1, \ldots, J$ is the region level such that $\sum_{j=1}^{J} n_j = k$. Let $\mathbf{X} \in \mathbb{R}^{k \times P}$ be the design matrix and $\gamma_\tau \in \mathbb{R}^{1 \times P}$ the unknown fixed parameter of interest where $\tau \in [0, 1]$ indicates the quantile level. The mixed quantile model [11] estimates the τ-quantile of the conditional distribution function of $\tilde{\mathbf{z}}$ as

$$Q_{\tilde{\mathbf{z}}|\mathbf{u}_\tau}(\tau \mid \mathbf{X}, \mathbf{u}_\tau) = \mathbf{X}\gamma_\tau^\top + \mathbf{u}_\tau$$

where $\mathbf{u}_\tau \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\tau)$ denotes the vector of region-specific random effects at τ level quantile. Weights at regional level (i.e., proportions of the Italian ECEC services in the j region) are inserted in the log-likelihood function.

Fixing $\tau = 0.75$, the estimated parameter for privates ECEC equals 0.25 while the public one 0.616 with p-values < 0.0001. After validating the effect of titularity, we now analyze the regional variability. Figure 4 shows the conditional quantile predictions with 0.95 confidence interval. We can note that some virtuous Italian region exist, i.e., Liguria, Trentino Alto Adige, Emilia-Romagna, and Lazio, where also private childcare services are defined as inclusive.

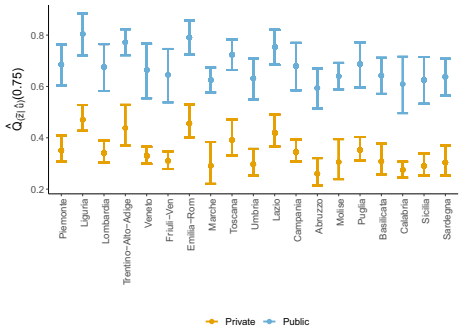


Fig. 4. Conditional quantile predictions at level $\tau = 0.75$ divided by type of ECEC services with 0.95 confidence interval where standard errors are calculated by bootstrap.

5 Discussion

The most important limitation in our work is that data used in our analysis can only capture partial aspects of inclusiveness, due to the use for which they were originally collected. Despite this, the use of advanced statistical methods can help to capture complex, multidimensional and not directly observable phenomena and to obtain useful information for policy makers at a finer geographical level. For example, private childcare services are not always characterized by less inclusiveness. Although this is generally true, in some regions private services are also virtuous. Moreover, the high heterogeneity of the observed phenomenon within the regions is also evident. This is a sign of the need for more uniform regulations on childcare services not only at a national but also at regional level.

References

1. Plantenga, J., Remery, C., Camilleri-Cassar, F.: The provision of childcare services: a comparative review of 30 European countries. Office for Official Publications of the European Communities, Luxembourg (2009)
2. Landivar, L.C., Ruppanner, L., Scarborough, W.J.: Are states created equal? Moving to a state with more expensive childcare reduces mothers' odds of employment. Demography **58**(2), 451–470 (2021)
3. Luci-Greulich, A., Thevenon, O.: The impact of family policies on fertility trends in developed countries: L'influence des politiques familiales sur les tendances de la fécondité des pays développés. Eur. J. Population **29**, 387–416 (2013)
4. Felfe, C., Lalive, R.: Does early child care affect children's development? J. Public Econ. **159**, 33–53 (2018)
5. Andreella, A., Aliverti, E., Caldura, F., & Campostrini, S. (2024). Spatial clusters for demand and supply of childcare services in Italy. arXiv:2401.07600 [stat.AP]
6. Andreella, A. & Campostrini, S. (2023). Suitable statistical approaches for novel policies: spatial clusters of childcare's services in Veneto, Italy. CARMA, 205–212. Editorial Universitat Politecnica de Valencia. https://doi.org/10.4995/CARMA2023.2023.16439

7. Dipartimento per le politiche della famiglia, ISTAT, Università Ca' Foscari, & MIPA (2020). Nidi e servizi educativi per la prima infanzia, stato dell'arte,criticità e sviluppi del sistema educativo integrato 0-6. https://istat.it/it/archivio/244158
8. Dipartimento per le politiche della famiglia, ISTAT, & Università Ca' Foscari (2022). I servizi educativi per l'infanzia in un'epoca di profondi cambiamenti. https://istat.it/it/archivio/287748
9. Darrell Bock, R., Lieberman, M.: Fitting a response model for n dichotomously scored items. Psychometrika **35**(2), 179–197 (1970)
10. Molina, I., Rao, J.: Small area estimation of poverty indicators. Can. J. Stat. **38**(3), 369–385 (2010)
11. Geraci, M., Bottai, M.: Linear quantile mixed models. Stat. Comput. **24**, 461–479 (2014)
12. Asparouhov, T.: Sampling weights in latent variable modeling. Struct. Equ. Model. **12**(3), 411–434 (2005)

Analyzing Compression of Comorbidity in Italy: A Mixture Model with Regional Random Effects

Angela Andreella[2(✉)] and Stefano Campostrini[1]

[1] Deparment of Economics, Ca' Foscari University of Venice, Venice, Italy
[2] Department of Economics and Management, University of Trento, Trento, Italy
angela.andreella@unitn.it
http://angeella.github.io

Abstract. This study pursues three objectives. Firstly, we propose a comorbidity index defined as a latent variable incorporating the disability weights from the Global Burden of Disease (GBD) projects into the estimation process. Secondly, we model the nonlinear relationship between this novel comorbidity index and socio-demographic features using a mixed-mixture model, which accommodates zero inflation and variability between the Italian regions. Lastly, an initial exploration of the concept of comorbidity compression in socio-demographic subpopulations is provided by analyzing the PASSI data across twelve years.

Keywords: Comorbidity · surveillance system PASSI · latent trait model · mixed beta ordered regression

1 Introduction

Comorbidity, the simultaneous occurrence of multiple diseases in one individual, is common, especially in older and chronically ill populations. It complicates diagnosis, treatment, and overall well-being. Recognizing comorbidity is vital for informed health decisions and resource allocation. The impact of multiple chronic diseases includes reduced quality of life, higher mortality, and increased healthcare costs. Analyzing comorbidity faces challenges, and understanding these complexities is crucial as the aging population grows [1,9].

The only viable approach to accommodate an aging population is by the lens of compressing morbidity, e.g., by preventing behavioral risk factors [8]. Introduced in 1980 as a counterpoint to the "Failures of Success" paradigm, the compression of morbidity theory says that the age of onset for aging and chronic diseases can be delayed more rapidly than the extension of life expectancy [5].

Despite Pastore et al. [8], Stival et al. [11] and Demeru and Virginia [4] works, Italy's comorbidity compression still needs to be explored. Pastore et al. [8] limited the analysis defining comorbidity as the presence of at least one chronic disease, overlooking the coexistence of multiple diseases and their impact on

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 27–33, 2025.
https://doi.org/10.1007/978-3-031-64431-3_5

an individual's quality of life. Instead, Stival et al. analyzed each disease separately, focusing on their geographical variability in the older population. They analyzed data from the Italian PASSI surveillance system [2] as in this work. In contrast, Demeru and Egidi [4] used data from the ISTAT Health Interview survey without analyzing socio-demographic differences. Our study addresses these gaps, introducing a latent model that estimates comorbidity using the disability weights from the Global Burden of Disease (GBD) project following Andreella et al. [1]'s analysis. However, in this work [1], the comorbidity was analyzed only for the year 2019, while, here, we focus on analyzing the compression of comorbidity using 12-year data from the Italian PASSI surveillance system [2]. Thanks to the latent trait model results, we then fitted a mixed-mixture model, which enables a first analysis of comorbidity compression within population subgroups defined by various socio-demographic variables.

2 Data

Our data stems from the Italian surveillance system PASSI, a cross-sectional study to recollect data about lifestyles and behavioral risk factors linked to the prevalence of Non-Communicable Diseases (NCDs). PASSI targets Italian adults aged 18 to 69 years. For more insight, refer to [2]. Additional information is available on the following webpage: https://www.epicentro.iss.it/passi/en/english.

The survey includes a question where respondents self-report the following NCDs: diabetes, kidney failure, bronchitis/emphysema/respiratory failure, myocardial infarction/cardiac ischemia/coronary artery disease, and tumors (including leukemias and lymphomas). We also analyze socio-economic variables such as sex (female/male), age (18 to 69), educational level (low if below high school, high otherwise), economic status (high if easy ends meet at the end of the month, low otherwise), along with the survey year (2008 to 2019) and region. The dataset comprises 53.18% female respondents, 60.17% people with higher education, and 52.77% respondents with some economic problems. We have 445, 709 observations in total. Some descriptive analyses are presented in [1] focusing on 2019.

3 Latent Trait Model

Fries and coauthors [5] emphasize that a measure of morbidity should focus on a broad form of disability, often using functional disability as a proxy. Reliably assessed by patient self-report, this metric aligns with patient perceptions of ill-health [5]. Following this perspective, we conceptualize comorbidity as a latent variable based on the concurrent presence of the 5 NCDs detected by PASSI. To enrich our definition of comorbidity in terms of functional disability and impact on the quality of life, we integrate the index proposed by Andreella et al. [1] based on the disability weights estimated from the GBD project [10]. These weights quantify the magnitude of health loss associated with specific health conditions.

Let consider the manifest variables X_i with $1 \leq i \leq 5$, i.e., the five NCDs detected by PASSI, distributed as a Bernoulli random variable with expected

value equals $\pi_i(\mathbf{z})$. The latent trait model [7] is then defined as:

$$\text{logit}\{\pi_i(\mathbf{z})\} = \beta_{0i} + \beta_{1i}\mathbf{z}$$

where β_{0i} can be interpreted as the prevalance of disease i and β_{1i} the effect of disease i in the comorbidity definition $\mathbf{z}$.

In the likelihood-based estimation process of the parameters β_{0i}, β_{1i} and $\mathbf{z}$, the observations $j = 1, \ldots, N$ are weighted as follow:

$$w_j = \frac{(D_j + 1)N}{\sum_{j=1}^{N} D_j + N}$$

where D_j is the comorbidity index proposed by Andreella et al. [1] using the multiplicative approach. This index is based on the disease weights estimated by the GBD project introduced before [10].

3.1 Results

Table 1 shows the estimated values of the $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ vector of coefficient parameters for each NCD detected by PASSI. Figure 1 represented the estimated probability of having the i disease across several $\hat{\mathbf{z}}$ (i.e., comorbidity) values. We can note that the less prevalent NCD is kidney failure, which is also the most impactful one in terms of morbidity. In contrast, tumors are the most prevalent and the least impactful. One possible explanation is that kidney problems pave the way for the onset of various disabling diseases. Conversely, cancers, while widespread, exhibit a spectrum of impact, ranging from mildly disabling to significantly disabling. However, the latter category is often influenced by a mortality effect (i.e., selecting respondents).

Table 1. Estimate values of the latent trait model's parameters.

Disease	$-\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$
Diabetes	3.611	1.378
Kidney	6.036	2.046
Bronchitis	3.373	1.134
Ima	4.819	1.820
Tumor	3.185	0.820

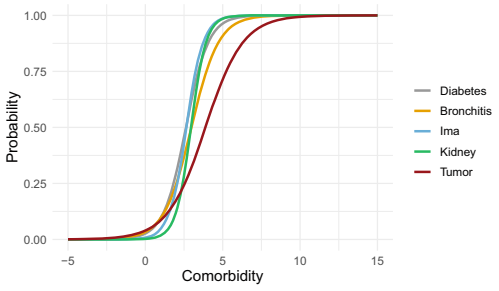


Fig. 1. Estimated probabilities to have the diseases $i = 1, \ldots, 5$ across several comorbidity values.

4 Ordered Beta Regression with Random Effect

After estimating the comorbidity index, we introduce in this section a model designed to analyze comorbidity compression within socio-demographic subpopulations. The objective is to leverage and encompass the health variability among Italian regions and take into account the particular distribution of the estimated comorbidity index, such as the presence of zero inflation.

Let define with $\mathbf{y}$ the estimated comorbidity index measured on the N respondents, i.e., $\mathbf{y} = (y_1, \dots, y_j, \dots, y_N)^\top$ which was rescaled to take values in $[0, 1]$, i.e., $y_j = (z_j - \min_{1 \le j \le N} z_j)/(\max_{1 \le j \le N} z_j - \min_{1 \le j \le N} z_j)$. Since the observations have a clustered structure defined by the geographical region, we define the region by $k = 1, \dots, K$ and the observation within the region as $l = 1, \dots, n_k$ such that $\sum_{k=1}^{K} n_k = N$ with K being the total number of regions (i.e., $K = 19$). The ordered beta regression [6] with random effect is estimated where we assume the following distribution for $\mathbf{y}$:

$$\begin{aligned}
\Pr(y_{lk} = 0) =& 1 - \text{logit}^{-1}(X_{lk}\boldsymbol{\beta}_0 - k_1) \\
\Pr(y_{lk} = 1) =& \text{logit}^{-1}(X_{lk}\boldsymbol{\beta} - k_2) \\
\Pr(0 < y_{lk} < 1) =& [\text{logit}^{-1}(X_{lk}\boldsymbol{\beta} - k_1 + u_k) - \text{logit}^{-1}(X_{lk}\boldsymbol{\beta} - k_2 + u_k)] \\
& \text{Beta}(\text{logit}^{-1}(X_{lk}\boldsymbol{\beta} + u_k), \psi)
\end{aligned}$$

where $\text{Beta}(\text{logit}^{-1}(X_{lk}\boldsymbol{\beta} + u_k), \psi)$ is the Beta distribution with mean parameter $\text{logit}^{-1}(X_{lk}\boldsymbol{\beta} + u_k)$ and precision parameter ψ.

The same set of fixed effects $\mathbf{X}\boldsymbol{\beta}$ enters all three probabilities definitions (i.e., age, sex, educational level, economic problem, year, and interaction between sex and age). Focusing on the probability to be $\mathbf{y} \in (0, 1)$, we insert a regional random effect $u_k \sim \mathcal{N}(0, \sigma_u^2)$ along with the fixed effects defined by $\mathbf{X}\boldsymbol{\beta}$.

The scalar parameter k_1 and k_2 are ordered cutpoints, i.e., $k_1 < k_2$ that permit to relate the three probabilities [6] and the probabilities are no longer exchangeable as in a basic zero-one inflated model. Kubinec [6] proposed a Bayesian estimation approach using an induced Dirichlet prior over the set of cutpoints $\{k_1, k_2\}$. However, we rely on the `glmmTMB` package [3], in which a MLE approach is used, assigning $k_1 = -1$ and $k_2 = 1$ as starting values.

4.1 Results

Figures 2 and 3 show the estimated comorbidity for subpopulations defined by different education and economic status levels for female and male respondents, respectively. Males have greater levels of comorbidity index than females in older ages, while the situation reverses in younger ages. In fact, we found a significant interaction between sex and age, giving a cue to a possible crossover effect [8].

Favorable social conditions, characterized by the absence of economic difficulties and a high level of education, were associated with a reduction in comorbidity levels, as evidenced by [1,8]. Specifically, our model predicts a comorbidity

level of 0.158 for males and 0.111 for females at age 69 in 2008 under favorable social circumstances. In contrast, when examining the most vulnerable subpopulation, characterized by low educational attainment and economic challenges, the predicted comorbidity levels are higher, reaching 0.234 for males and 0.174 for females, again in 2008 at age 69. Encouragingly, the phenomenon of the compression of morbidity is observed not only in socially advantaged groups but also in high-risk populations and those facing social disadvantages.

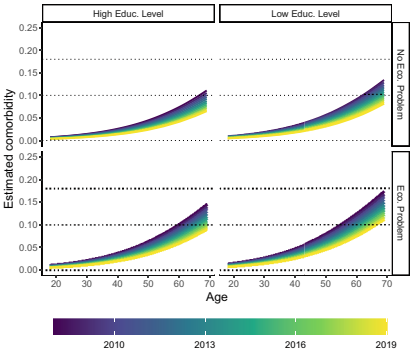


Fig. 2. $\hat{y}$ across age for the **female** subpopulation within each group categorized by levels of education and economic status. Colors denote the survey year from 2008 to 2019.

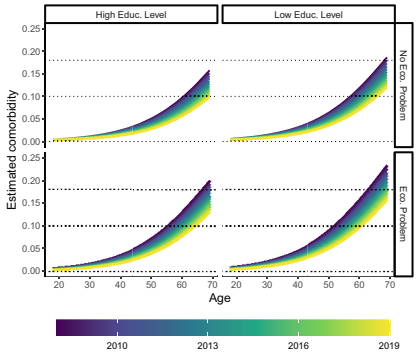


Fig. 3. $\hat{y}$ across age for the **male** subpopulation within each group categorized by levels of education and economic status. Colors denote the survey year from 2008 to 2019.

Finally, Fig. 4 presents the Best Linear Unbiased Predictions (BLUP) for each Italian region. Notably, the data highlights that the Umbria region exhibits the most challenging scenario in terms of comorbidity. Overall, the southern regions demonstrate a higher level of comorbidity. The results are not surprising since these regions present a worse profile in terms of risk factors behaviors.

5 Discussion

This paper introduces a comorbidity measure obtained from NCDs identified through the PASSI surveillance system and disability weights estimated by the GBD project. The weighted latent variable is examined using a mixed-mixture model that considers regional variability. Through these analyses, we can confidently assert a positive outcome: comorbidity compression is acting. In addition, applying the model in subpopulations based on low, medium, and high levels of behavioral risk factors provides similar coefficient values related to the year effect, suggesting that compression is not solely due to reduced risk factor prevalence. Further analysis is needed for precise conclusions, and future work will compare results with official data for a comprehensive understanding.

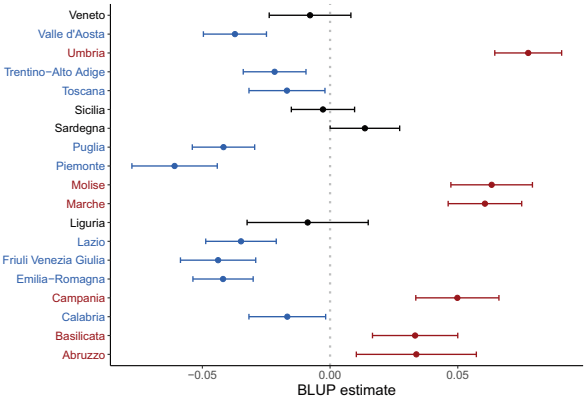


Fig. 4. Best linear unbiased prediction (BLUP)

Acknowledgments. Angela Andreella gratefully acknowledges funding from the PON 2014–2020/DM 1062 grant of the Ca' Foscari University of Venice, Italy.

References

1. 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**(1), 18 (2023)
2. Baldissera, S., et al.: Peer reviewed: features and initial assessment of the Italian behavioral risk factor surveillance system (PASSI). Preventing Chronic Dis. **8**(1) (2011)
3. Brooks, M.E., Kristensen, K., Van Benthem, K.J., Magnusson, A., Berg, C.W., Bolker, B.M.: . glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. **9**(2), 378–400 (2017)
4. Demuru, E., Egidi, V.: Adjusting prospective old-age thresholds by health status: empirical findings and implications. a case study of Italy. Vienna Yearb. Popul. Res. **14**, 131–154 (2016)
5. Fries, J.F., Bruce, B., Chakravarty, E.: Compression of morbidity 1980–2011: a focused review of paradigms and progress. J. Aging Res. **2011**(1), 261702 (2011)
6. Kubinec, R.: Ordered beta regression: a parsimonious, well-fitting model for continuous data with lower and upper bounds. Polit. Anal. **31**(4), 519–536 (2023)
7. Moustaki, I., Knott, M.: Generalized latent trait models. Psychometrika **65**, 391–411 (2000)
8. Pastore, A., Tonellato, S.F., Aliverti, E., Campostrini, S.: When does morbidity start? An analysis of changes in morbidity between 2013 and 2019 in Italy. Stat. Methods Appl. **32**(2), 577–591 (2023)

9. Prados-Torres, A., Calderón-Larrañaga, A., Hancco-Saavedra, J., Poblador-Plou, B., van den Akker, M.: Multimorbidity patterns: a systematic review. J. Clin. Epidemiol. **67**(4), 254–266 (2014)
10. Salomon, J.A., et al.: Disability weights for the global burden of disease 2013 study. Lancet Glob. Health **3**(11), 712–723 (2001)
11. 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 (2024) [stat.AP]



Beyond Correlation: Discovering Causal Effects of AI on Innovation Across European SMEs

Lea Anna Cozzucoli(✉) and Francesco Pauli

Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy
leaanna.cozzucoli@phd.units.it, francesco.pauli@deams.units.it

Abstract. Nowadays, causal inference methods using information from observational studies play a fundamental role in empirical research, in order to understand the effects of interventions, treatments and policies in different fields. This work aims to measure the causal effect of Artificial Intelligence (AI) and Big Data Analytics (BDA) on innovation in businesses accounting for heterogeneous causal effects within the potential outcomes framework, with a particular focus on the conditional average treatment effect (CATE). The data analyzed in the study are taken from a survey on European Small and Medium Enterprises (SMEs). Propensity score matching (PSM) is used to address possible imbalances in the sample. Afterwards, CATE is estimated using Bayesian Additive Regression Trees (BART). Our results suggested a significant and varying effect across different enterprise sizes, highlighting the importance of considering heterogeneity in causal effects for effective policy interventions.

Keywords: Causality · Digitalization · Heterogeneity · Observational studies

1 Introduction

Causal inference is a crucial part of empirical research, allowing scientists and policymakers to determine the effects of different interventions, treatments, and policies in many fields. In the potential outcome framework, causal effects are measured by comparing the observed outcome and the counterfactual outcome that would have been observed under a different treatment [1]. Traditionally, in ideal cases, the effect of the treatment on an outcome is inferred from randomized controlled trials (RCTs), which are considered the gold standard for establishing causality. However, such trials are not always practical or even ethical. Consequently, developing statistical methods and frameworks for causal inference in observational data has become crucial. Differently from the RCTs, where no interference between the outcome and the treatment is present, the main challenge in observational studies is to differentiate between bare associations and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 34–39, 2025.
https://doi.org/10.1007/978-3-031-64431-3_6

causal relationships that are essential for understanding mechanisms and making predictions for interventions.

In the last decades, causal inference methods have been largely developed, and, under some assumptions, it is possible to evaluate the causal effect, better known as a true causal effect, even in non-experimental and quasi-experimental studies. However, a fundamental challenge arises in the presence of variability in causal effects among different individuals or groups within the population. Those differences are related to other baseline characteristics of individuals or groups, both measurable and unmeasurable. The concept of heterogeneity in causal effects is fundamental since recognizing and analyzing such heterogeneity is essential to find interventions or policies with maximum effectiveness that, at the same time, can reduce and mitigate possible disparities. A typical way for measuring treatment effect in the presence of heterogeneity is to estimate the conditional average treatment effect (CATE) [2].

In this work, we aim to measure the heterogeneous causal effect, in terms of CATE, for a binary treatment and outcomes under the potential outcomes framework. The main idea is to measure the causal effect of Artificial Intelligence (AI) and Big Data Analytics (BDA) on innovation introduction in business.

2 Measuring Heterogeneous Causal Effect

The true causal effect represents the actual effect of a treatment on a potential outcome. It follows that, to measure causal effects, no influence from other factors except the treatment should be present; this means that there must be no difference between the group of treated units and the group of not treated (control) units in terms of baseline characteristics. In terms of probability, this can be written

$$\mathbb{P}(\mathbf{X}, \mathbf{U}|Z = 1) = \mathbb{P}(\mathbf{X}, \mathbf{U}|Z = 0), \tag{1}$$

where $\mathbf{X}$ and $\mathbf{U}$ represent measurable and unmesurable covariate matrices while Z is the treatment variable. For simplicity, the treatment variable will be assumed as dichotomous, $Z \in \{0, 1\}$. This first statement guarantees the so-known unconfundness of the treatment assignment, meaning that $\mathbf{X}, \mathbf{U} \perp\!\!\!\perp Z$ and, is completely satisfied in RCTs. Other strong assumptions necessary to exploit causal inference methods are the overlap and the Stable Unit Treatment Value Assumption (SUTVA). The first guarantees that both treated and untreated units have a non-zero probability of being allocated to either the treatment or control groups and, jointly with the assumption of unconfoundedness (1), define the strong ignorability. The SUTVA assumption imposes no interference between units, implying that potential outcomes for the i-th unit are unrelated to the treatment status of other individuals [3]. SUTVA is the key assumption in the causal inference framework necessary to define causal estimands. While the latter condition is more easily satisfied in observational studies the first two may sometimes be unsatisfied.

To measure the direct effect of a treatment, widely used measures are the average treatment effect (ATE) and the average treatment effect for the treated

(ATT). Under the assumption of strong ignorability and SUTVA the ATE is defined as

$$\mathbb{E}\left[Y_i(Z=1) - Y_i(Z=0)\right]. \tag{2}$$

It enables to estimate the average difference in outcomes between treated and controls. The lack of validity of the assumptions of unconfoundedness, overlap and the SUTVA can lead to biases in the previously proposed estimator and, thus, a sort of pre-treatment of the data must be done to ensure that all the assumptions are satisfied. To remove the effect of confounders while estimating the causal effect of a treatment over the outcome, propensity score matching methods (PMS) are widely employed. The propensity score (PS) is defined as the probability of treatment assignment conditional on observed baseline covariates. For the i-th unit is defined as

$$e_i(\mathbf{X}) = \mathbb{P}(Z_i = 1|X_i). \tag{3}$$

Since in observational studies, the PS for each unit is not known a priori, it is estimated from the data using regression models (commonly logistic regression). Once the value is estimated, it is used to achieve the covariate balance that ensures that the assumption of unconfundness holds. [4] showed two relevant properties of the PS:

$$\begin{aligned} &Z_i \perp\!\!\!\perp X_i|e(X_i) \\ &Y_i(Z=1), Y_i(Z=0) \perp\!\!\!\perp Z_i|e(X_i), \end{aligned} \tag{4}$$

where the first is known as balancing property. Specifically, PMS is a non-parametric pre-processing procedure that helps reduce the parametric model dependence while limiting the sample dimension and forming matching sets of treated and control units that share similar values of the PS. A possible limitation of the method is that the balance is achieved only on the covariates used in the PS calculation (3), so, it should hold that there are no unmeasured characteristics, which is not always true. The above-cited method are quite useful while dealing with causal effects in observational data where not all the mentioned conditions are satisfied. However, there are cases in which taking into account the data heterogeneity can be even more interesting than estimating the "classical" causal effect. While trying to do so, a common alternative could be to provide information about a causal effect with respect to some covariates using, for example, the Conditional Average Treatment Effect (CATE). The estimator enables us to estimate how the response to a treatment varies with respect to some baseline characteristics of the units and it can be seen as a conditional ATE.

$$\mathbb{E}\left[Y_i(Z=1) - Y_i(Z=0)|X=x\right]. \tag{5}$$

3 Application

In this section, we propose an application of causal inference methods in the presence of heterogeneous effects considering some observational data. Specifically, the aim is to measure the effect on innovation of newly developed methods in data analysis, such as artificial intelligence methods, in enterprises.

3.1 Data

The data were taken from a survey conducted by Gesis on behalf of the European Commission in 2020: "*Flash Eurobarometer 486*" [5]. The survey aims to collect information about sustainability practices, digitalization, and growth, in European Small and Medium Enterprises (SMEs'), which are largely diffused in Europe. In this case, the treatment variable was defined as using certain methods such as Artificial Intelligence (e.g. machine learning) and/or Big data analytics (e.g., data mining); so the enterprises that state to use one or both technologies were taken as treated units. The effect was measured over the introduction of innovation in the business during the past 12 months from the interview, i.e. new or significantly improved product, service, process etc. Furthermore, the dataset provided a range of variables that were considered as a source of heterogeneity within the SMEs. These variables include the activity year, enterprise dimension, turnover, turnover increase, and the enterprise's location. Further details such as the industrial area, the type of company, export activities, ownership structure, affiliation with national or international groups, and whether the enterprise is family-owned, usage of other digital technologies and implementation of sustainability procedures were also considered.

3.2 Analysis

As previously stated, the aim is to estimate the causal effect of the employment of AI and/or BDA instruments on innovation development. Since we are dealing with observational data, we first check for potential imbalances in order to subsequently attempt propensity score matching (PMS) using full matching (for more details refer to [6]). In the analysis, a potential population of 11,072 SMEs is considered. The sample size remains the same after full matching, since the method aims to retain all units by optimally pairing treated and control groups. For all subsequent preliminary analyses related to PMS, the R package `matchit` [7] is used. To achieve matching it is first necessary to assess the propensity score for each unit using baseline characteristics. In this case, logistic regression on AI and BDA use with respect to pre-treatment covariates previously cited is used.

As reported in Fig. 1, before matching, the data were not balanced majorly in terms of dimension, turnover, export, and use of other digital technologies. The matched sample is used in subsequent analysis to guarantee more robust results.

At this point, we were interested in measuring the ATE conditional to a specific variable that, in this case, will be the dimension of the enterprise distinguishing between micro, small, or medium companies. The hypothesis is that the effect would vary with respect to the chosen covariates. Different methods have been proposed in literature to estimate the CATE. For this work, we relied on the Bayesian Additive Regression Tree (BART) introduced by [8]. All the analyses were performed using the `R` package `bartCause` [9]. In Fig. 2, the results suggest a slight difference in terms of ATE in each group; specifically, the CATE for the micro-enterprises is 0.1627 (CI: 0.1305 - 0.1950) while for the medium it is 0.1314 (CI:0.09-0.1669). The overall ATE is 0.1535 (CI: 0.1257-0.1812).

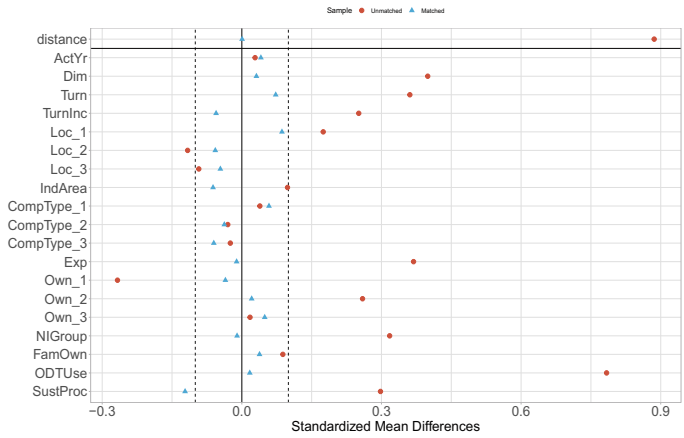


Fig. 1. Comparison in terms of standardized mean difference of covariate balance before and after propensity score matching.

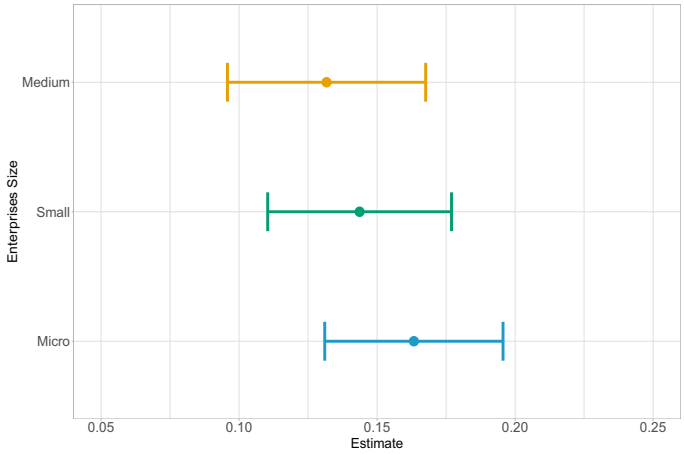


Fig. 2. Estimated CATE across SME Size Categories. The figure shows point estimates and 95% credible intervals for the CATE.

4 Discussion

The results obtained suggests that the implementation of AI and/or BDA has a significant effect on business innovation. Specifically, ATE estimates highlighted the positive overall effect, on average, of AI and BDA on innovation activities in businesses. The CATE estimation shows limited heterogeneity of causal effects, with a slightly greater effect in smaller companies. Micro-enterprises seem to benefit more from adopting these technologies compared to medium enterprises, as reflected by their higher CATE value. The varying degrees of effectiveness based

on the size of the enterprise highlights the importance of tailored approaches in technology adoption and policy-making to foster innovation effectively. Consequently, the results suggest that digital technologies can be a key driver to innovation in enterprises.

We believe there is potential for further developments based on our findings, from both methodological and practical point of view. Firstly, our goal is to give more emphasis on variable selection in matching process and also assess the benefits of other possible estimators for heterogeneous causal effect. From a practical perspective, various factors should be evaluated and included, along with investigating the long-term impacts of adopting AI and Big Data Analytics on innovation.

References

1. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press (2015)
2. Wang, J., Liu, C.: Estimation of conditional average treatment effect by covariates balance methods. Stat. Interface **15**(3), 312–322 (2022)
3. 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)
4. Rosenbaum, R.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
5. European Commission.: Flash eurobarometer 486: Smes, start-ups, scale-ups and entrepreneurship (2020)
6. Hansen, B.B., Klopfer, S.O.: Optimal full matching and related designs via network flows. J. Comput. Graph. Stat. **15**(3), 609–627 (2006)
7. Ho, D.E., Imai, K., King, G., Stuart, E.A.: MatchIt: nonparametric preprocessing for parametric causal inference. J. Stat. Softw. **42**(8), 1–28 (2011)
8. Hill, J., Linero, A., Murray, J.: Bayesian additive regression trees: a review and look forward. Annu. Rev. Stat. Appl. **7**(1), 251–278 (2020)
9. Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. **20**(1), 217–240 (2011)



Simultaneous Clustering and Dimension Reduction of Functional Data

Roberto Rocci[1] and S. Antonio Gattone[2(✉)]

[1] Sapienza University, Rome, Italy
rocci@uniroma1.it
[2] G. d'Annunzio University, Pescara, Italy
gattone@unich.it

Abstract. In this work a technique that combines clustering and dimension reduction for functional data is proposed. The optimal subspace and partition are identified by optimizing a unique loss function. To prevent the identification of subspaces with very low deviance, a penalty equal to the negative total deviance in the reduced space is added to the objective function. It can be shown that by varying the value of the penalty, the proposed loss function could recover several existing methods such as the Functional Reduced K-means and the Functional Factorial K-means. Finally, to consider the functional nature of the data, regularization is also incorporated in the estimation.

Keywords: Functional Data Analysis · Clustering · Dimension reduction · Smoothing

1 Introduction

Functional data analysis (FDA) deals with the analysis and theory of data that are in the form of functions whose measurements vary along a continuum [8]. Functional data arises in a variety of domains such as environmental monitoring, medical imaging, chemistry, biology, etc. [10]. Standard multivariate techniques such as regression, dimension reduction, clustering, classification, need to be suitably modified to cope with the high dimensionality due to the continuous nature of measurements [8,11].

Functional clustering aims at identifying heterogeneous pattern within continuous functions. Literature on clustering methods for functional data is indeed extensive and we refer the readers to review papers such as [5,6,16]. One common approach to cope with the infinite-dimensional nature of the data space is to use basis functions to represent each function. With this finite-dimensional representation, standard clustering techniques, such as K-means, can be applied to the coefficients of the basis functions [1,3].

On the other hand, Functional Principal Component Analysis (FPCA) is commonly used to obtain a lower dimensional representation while capturing the essential features of the functional data [9].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 40–45, 2025.
https://doi.org/10.1007/978-3-031-64431-3_7

Some authors [2,7,12] combine clustering and dimension reduction by applying clustering algorithms on the functional principal component scores. This approach, known for multivariate data as *tandem analysis*, has the disadvantage that the main modes of variation identified by the PC scores may not contribute to the identification of the cluster structure. This problem has been addressed in the FDA framework by incorporating dimension reduction and clustering in a single objective function: the functional reduced K-means (FRKM) [4,13], the functional factorial K-means (FFKM) [15], and the functional generalized reduced clustering (FGRC) [14]. In particular, FFKM aims to minimize the within-cluster deviance in the reduced space. FRKM, on the other hand, maximizes the between-cluster deviance in the reduced space. FGRC identifies subspaces for clusters and subspaces for a disturbing structure. The FGRC loss function involves two tuning parameters, which are calibrated using the Gap statistic.

In this work we start from a limitation of FFKM: it may identify dimensions with negligible total deviance leading to insignificant partitions. To overcome this weakness, we introduce a penalty equal to the negative total deviance in the reduced space. Interestingly, FFKM and FRKM can be recovered by tuning the penalty. All clustering methods for functional data require transforming raw data into smooth functions. However in the clustering FDA literature, this step is performed *a priori* and independently of the subsequent cluster analysis. This can lead to confusion between measurement error and true clustering structure. Following [4], we add a regularization term to the loss function which smooths the estimates of cluster centroids improving estimation efficiency and interpretability. In this way, smoothing is integrated with subspace and cluster identification.

In the next section the proposed clustering method is presented through the derivation of the loss function and its regularized version.

2 The Method

We first introduce the two main tools of the tandem approach, *i.e.* the functional K-means (Subsect. 2.1) and the FPCA (Subsect. 2.2). Subsection 2.3 introduces the new clustering method for functional data, named Functional Projection K-means (FPKM). The regularized version is presented in Subsect. 2.4.

2.1 Functional K-Means

Let $\{f_i(t)\}_{i=1}^I = \{f_1(t), f_2(t), \ldots, f_I(t)\}$ be a set of observed curves where $t \in \mathcal{T}$ and $\mathcal{T}$ is an interval of $\Re$. Such data functions lie in a separable Hilbert space H with inner product $\langle f, g\rangle$ and norm $\|f\|^2 = \langle f, f\rangle$.

The goal is to cluster the sample of curves $\{f_i(t)\}_{i=1}^I$ into G homogeneous groups. The functional K-means (FKM) assumes the data arising from the following model

$$f_i(t) = m_g(t) + \epsilon_i(t) \qquad i = 1, \ldots, I; g = 1, \ldots, G; \tag{1}$$

where m_g are smooth unknown centroid functions and $\epsilon_i(t)$ denote an unobservable zero-mean stochastic error term. The FKM loss function to be minimized is

$$L_{\text{FKM}}(\theta) = \sum_{ig} u_{ig} \left\| f_i - m_g \right\|^2 \tag{2}$$

where $\theta = \{u_{11}, \ldots, u_{IG}, m_1, \ldots, m_G\}$ and $u_{ig} \in \{0, 1\}$, $\sum_g u_{ig} = 1 \ \forall i$ with $u_{ig} = 1$ if $f_i(t)$ belongs to the g-th cluster and $u_{ig} = 0$ otherwise.

2.2 FPCA

Functional principal component analysis (FPCA) is the main tool of dimension reduction in FDA. FPCA projects the functional data onto a reduced space spanned by a few principal component functions. The underlying model behind FPCA can be defined as follows

$$f_i(t) = \sum_{q=1}^{Q} c_{iq}\psi_q(t) + \epsilon_i(t), \qquad i = 1, \ldots, I; \tag{3}$$

where $\psi_q(t)$ are smooth unknown orthonormal principal functions, *i.e.* $\langle \psi_q, \psi_r \rangle = 1$ if $q = r$ and $\langle \psi_q, \psi_r \rangle = 0$ if $q \neq r$. The loss function minimized by FPCA is

$$L_{\text{FPCA}}(\theta) = \sum_i \left\| f_i - \sum_q c_{iq}\psi_q \right\|^2 \tag{4}$$

where $\theta = \{c_{11}, \ldots, c_{IQ}, \psi_1, \ldots, \psi_Q\}$. Let $f_i^p(t) = \sum_q c_{iq}\psi_q(t)$ be the orthogonal projection of $f_i(t)$ on the functional low-dimensional subspace found by FPCA, the tandem analysis estimates the data partition by minimizing

$$L_{\text{TA}}(\theta) = \sum_{ig} u_{ig} \left\| f_i^p - m_g \right\|^2 \tag{5}$$

where $\theta = \{u_{11}, \ldots, u_{IG}, m_1, \ldots, m_G\}$.

2.3 Functional Projection *K*-Means

The subspace spanned by the principal functions is optimal in reproducing the data deviance but not necessarily the clustering structure. This drawback of tandem analysis can be avoided by estimating simultaneously the subspace and the partition. In other terms, we have to to find a low-dimensional subspace spanned by the basis functions $\{\psi_q(t)\}_{q=1}^{Q} = \{\psi_1(t), \ldots, \psi_Q(t)\}$, say H_Ψ, containing the most relevant clustering information. To this end, we assume the projections to be represented in the following setup

$$f_i^p(t) = m_g(t) + \epsilon_i(t) \qquad i = 1, ..., I; g = 1, ..., G. \tag{6}$$

The FFKM [15] simultaneously finds an optimal projection of the curves $\{f_i(t)\}_{i=1}^{I}$ and a partition minimizing the within-deviance in the reduced space, *i.e.* by minimizing the loss

$$L_{\text{FFKM}}(\theta) = \sum_{i=1}^{I}\sum_{g=1}^{G} u_{ig} \left\| f_i^p - m_g \right\|^2 \tag{7}$$

where $\theta = \{u_{11}, \ldots, u_{IG}, c_{11}, \ldots, c_{IQ}, \psi_1, \ldots, \psi_Q, m_1, \ldots, m_G\}$.

It is important to note that the minimization of (7) implies that the centroids $m_g(t)$ lie in the subspace H_Ψ. Because of this, from now on, the centroids will be of the form

$$m_g(t) = \sum_{q=1}^{Q} a_{gq}\psi_q(t). \tag{8}$$

A potential drawback of FFKM is that it could find a solution whose projections have a trivial deviance that makes the identified partition insignificant. In order to prevent the aforementioned problem, we propose to modify the objective function (7) by introducing a penalty equal to the negative total deviance in the reduced space. The loss function of this new clustering method, called functional projection K-means (FPKM), is given by

$$L_{\text{FPKM}}(\theta; \alpha) = \sum_{i=1}^{I}\sum_{g=1}^{G} u_{ig} \left\| f_i^p - m_g \right\|^2 - \alpha \sum_{i=1}^{I} \left\| f_i^p \right\|^2 \tag{9}$$

where $\theta = \{u_{11}, \ldots, u_{IG}, c_{11}, \ldots, c_{IQ}, \psi_1, \ldots, \psi_Q, a_{11}, \ldots, a_{GQ}, \}$.

The second term in equation (9) penalizes low variance subspaces. The tuning parameter α varies from 0 to ∞. Interestingly we can show that by varying α the FPKM loss encompasses a wider class of existing methods. FFKM is obtained with $\alpha = 0$ and $G < I$. With $\alpha = 1$ and $G < I$, the loss in (9) corresponds to another popular method that simultaneously performs dimension reduction and clustering: the FRKM. The FRKM objective function is given by

$$L_{\text{FRKM}}(\theta) = \sum_{i=1}^{I}\sum_{g=1}^{G} u_{ig} \left\| f_i - m_g \right\|^2 \tag{10}$$

where $\theta = \{u_{11}, \ldots, u_{IG}, a_{gq}, \ldots, a_{GQ}, \psi_1, \ldots, \psi_q\}$. It can be shown that (10) is equivalent to (9) when $\alpha = 1$.

Finally, we can note that when $\alpha \neq 0$ and $G = I$ the objective function in (9) corresponds to FPCA, while, when $Q = I$, it corresponds to FKM.

2.4 The Regularized Functional Projection K-means

Following the spirit of functional data analysis, we assume that the underlying process generating the centroids is a smooth continuous curve. The smoothing

step is incorporated in the analysis by adding a regularization term to criterion (9). The loss function of the regularized FPKM is given by

$$L_{FPKM}(\theta;\alpha,\lambda) = \sum_{i=1}^{I}\sum_{g=1}^{G} u_{ig}\left\|f_i^p - m_g\right\|^2 - \alpha\sum_{i=1}^{I}\left\|f_i^p\right\|^2 + \lambda\sum_{g=1}^{G}\int\left[m_g''\right]^2 dt. \tag{11}$$

where $m_g''(t)$ is the second derivative of $m_g(t)$. The tuning parameter λ controls the roughness of the centroids, with larger λ values producing smoother centroids.

By varying the degree of smoothness and the level of subspace deviance different cluster solutions can be obtained so that a key factor is the proper selection of the tuning parameters α and λ. In the extended version of this work we will investigate an alternating least squares algorithm to estimate the model parameters and two data-driven procedures in order to automatically calibrate the two tuning parameters. The effectiveness of the proposal will also be tested on real and simulated data.

References

1. Abraham, C., Cornillon, P., Matzner-Løber, E., Molinari, N.: Unsupervised curve clustering using b-splines. Scand. J. Stat. **30**(3), 581–595 (2003)
2. Chiou, J.M., Li, P.L.: Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society. Series B **69**(4), 679–699 (2007)
3. Denis, C., Lebarbier, E., Lévy-Leduc, C., Martin, O., Sansonnet, L.: A Novel Regularized Approach for Functional Data Clustering: An Application to Milking Kinetics in Dairy Goats. J. R. Stat. Soc.: Ser. C: Appl. Stat. **69**(3), 623–640 (2020)
4. Gattone, S., Rocci, R.: Clustering curves on a reduced subspace. J. Comput. Graph. Stat. **21**, 361–379 (2012)
5. Hitchcock, D., Greenwood, M.: Clustering functional data. In: Hennig, C., Meila, M., Rocci, R., Murtagh, F., (eds.) Handbook of Cluster Analysis, chap. 13, pp. 265–288. Chapman and Hall/CRC (2015)
6. Jaques, J., Preda, C.: Functional data clustering: a survey. Adv. Data Anal. Classif. **8**, 231–255 (2014)
7. Peng, J., Müller, H.: Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. Annals of Applied Statistics **2**(3), 1056–1077 (2008)
8. Ramsay, J., Silvermann, B.: Functional Data Analysis. Springer-Verlag, New-York (2005)
9. Shang, H.: A survey of functional principal component analysis. Adv. Stat. Anal. **98**, 121–142 (2014)
10. Ullah, S., Finch, C.: Applications of functional data analysis: A systematic review. BMC Med. Res. Methodol. **13**(43), 1–12 (2013)
11. Wang, J.L., Chiou, J.M., Müller, H.G.: Functional data analysis. Annual Review of Statistics and Its Application **3**(1), 257–295 (2016)
12. Wu, R., Wang, B., Xu, A.: Functional data clustering using principal curve methods. Communications in Statistics - Theory and Methods **51**, 7264–7283 (2022)

13. Yamamoto, M.: Clustering of functional data in a low-dimensional subspace. Adv. Data Anal. Classif. **6**, 219–247 (2012)
14. Yamamoto, M., Hwang, H.: Dimension-reduced clustering of functional data via subspace separation. J. Classif. **34**, 294–326 (2017)
15. Yamamoto, M., Terada, Y.: Functional factorial k-means analysis. Comput. Stat. Data Anal. **79**, 133–148 (2014)
16. Zhang, M., Parnell, A.: Review of clustering methods for functional data. ACM Trans. Knowl. Discov. Data **17**(91), 1–34 (2023)



Comparing Methods for Testing Equality of Means in One-Way Designs

Rosa Arboretti[1], Elena Barzizza[2(✉)], and Nicolò Biasetton[2]

[1] University of Padova, Department of Civil Engineering, 35100 Padova, Italy
rosa.arbotetti@unipd.it
[2] University of Padova, Department of Management Engineering, 36100 Vicenza, Italy
elena.barzizza@phd.unipd.it, nicolo.biasetton@unipd.it

Abstract. This short paper addresses the problem of testing equality of means across multiple samples in one-way designs with unequal variances. Specifically, the focus is on assessing the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_S$ against ordered alternatives, as this problem is of practical interest. However, justifying the assumption of homogeneity of variances in practical applications presents a significant challenge. To handle with this problem, various solutions have been proposed in the literature. Recently, parametric bootstrap approaches have been introduced, demonstrating good performances, particularly for symmetric distributions. In this paper, we put forward a novel approach based on permutations, utilizing the Non-Parametric Combination (NPC) methodology. We aim to compare this new approach with the aforementioned methods in a real case study.

Keywords: Permutation · Case study · Non Parametric Combination Methodology

1 Introduction

A well debated statistical problem in the literature is the issue of making inference about the means of two or more populations without assuming equal variances or known variances ratio. Certainly, this issue emerges in numerous practical applications where justifying the homogeneity of variances among groups proves to be challenging and may result in inflated Type-I error [1,2,4,5,7,9,10]. Several solutions have been suggested, as exemplified by Wilcox (1987) [12] and Oshima et al. (1992) [8], as well as those outlined in Mondal et al. (2023) [6]. In this short paper we will focus on the problem of testing the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_S$ against the ordered alternatives in one-way design when the variances are unequal. Recently, Mondal et al. (2023) [6] introduced a parametric bootstrap procedure based on four different test statistics (a likelihood ratio test and a MinT test and their asymptotic counterpart) to deal with this problem. Nevertheless, it has been noted that these methods exhibit

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 46–52, 2025.
https://doi.org/10.1007/978-3-031-64431-3_8

robustness primarily in the case of symmetric distributions. We proposed an alternative approach based on permutation which involves the Non Parametric Combination methodology (NPC) [11]. Indeed this methodology presents some interesting advantages. Firstly, it is a permutation-based technique, implying that it does not necessitate any assumptions about the distribution. Assuming normality of the data is often unrealistic in practical applications, as it's unlikely for this distribution to occur in real-world scenarios [13]. Furthermore, the NPC-based approach potentially allows for obtaining an exact solution, provided that exchangeability assumption is justified under the null hypothesis and the cumulative distribution functions do not intersect under the alternative hypothesis. Formal proofs of these properties are in [11]. In addition, it has been noted that it exhibits strong power performance even with small sample sizes [3].

Section 2 is devoted to the explanation of the NPC-based methodology, in Sect. 3 introduces a case study where both approaches are applied. Final remarks are provided in Sect. 4.

2 Methodology

2.1 Parametric Bootstrap Approach

In this paper we are interested on testing the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_S \tag{1}$$

against the ordered alternatives:

$$H_1 : \mu_1 \leq \mu_2 \leq ... \leq \mu_S \tag{2}$$

with at least a strict inequality. Mondal et al. (2023) [6] proposed four test procedures: likelihood ratio test (LRT), an asymptotic version of likelihood ratio test (ALRT), a Min-T test heuristic procedure (Min-T) and its asymptotic version (AMin-T). In particular, a parametric bootstrap approach is used to assess the critical points of LRT, Min-T, and AMin-T. While, for ALRT the critical value is simply $\chi^2_{k-2,\alpha}$. A classical parametric bootstrap is employed, where for a given sample, the critical value is derived by sampling from an estimated normal distribution $N(0, S^2)$ under the null hypothesis and subsequently computing the test statistic. For a detailed explanation of the algorithms, please consult Mondal et al. (2023) [6] on pages 1919–1924.

2.2 NPC Methodology

In this section, we introduce an alternative permutation-based methodology to the solution proposed by Mondal et al. [6]. Let us consider our dataset X characterised by $S > 2$ sample, where $X = \{X_k, k = 1, \cdots, S\}$. We are interested on testing the null hypothesis 1 against the ordered alternatives hypothesis 2 with

at least a strict inequality. The basic idea is to split the dataset into two pooled groups $Z_{(1)m}$ and $Z_{(2)m}$:

$$Z_{(1)m} = X_1 \biguplus \cdots \biguplus X_m, Z_{(2)m} = X_{m+1} \biguplus \cdots \biguplus X_S \tag{3}$$

where $Z_{(1)m}$ is obtained by pooling the data from the first m groups and $Z_{(2)m}$ from the remaining data. The symbol $\biguplus$ denotes data pooling. In this way, the S-sample problem is transformed in a two-sample problem where the null hypothesis is written as $H_{0,m} : \mu'_1 = \mu'_2$ against $H_{1,m} : \mu'_1 \leq \mu'_2$ with μ'_1 and μ'_2 the observed mean respectively of $Z_{(1)m}$ and $Z_{(2)m}$. Consequently, the initial system of hypothesis reported in Eq. 1 and in Eq. 2 must be reformulated as:

$$H'_0 : \bigcap_{m=1}^{S-1} H_{0,m} \qquad H'_1 : \bigcup_{m=1}^{S-1} H_{1,m} \tag{4}$$

Now, let's turn our attention to the solution based on permutations for a generic two-sample problem. In an univariate setting, this solution corresponds to the first phase of the Non Parametric Combination methodology (see Pesarin et al. [11] for a detailed explanation). Be Y our dataset characterized by two samples, $Y_1 = Y_{1i}, i = 1, \cdots, n_1$ and $Y_2 = Y_{2i}, i = 1, \cdots, n_2$. Firstly we applied our selected test statistic T to our data Y, to obtain our observed test statistic $T^0 = T(Y)$. In particular, we adopted two different test statistics, the mean difference ($T^{\text{NPC}}_{\text{mean}}$) and the Welch t-test ($T^{\text{NPC}}_{\text{Welch}}$):

$$T^{\text{NPC}}_{\text{mean}} = \bar{Y}_2 - \bar{Y}_1, T^{\text{NPC}}_{\text{Welch}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{s^2_{Y_1}}{n_1} + \frac{s^2_{Y_2}}{n_2}}}$$

where, $\bar{Y}_1$, $\bar{Y}_2$, s_{Y_1} and s_{Y_2} are the sample means and their standard deviations of respectively the Y_1 and Y_2.

Then, we permuted randomly the dataset and we applied again our test statistic T to our permuted dataset Y^* obtaining our permuted test statistic $T^*_1 = T(Y^*)$. This random permutation must be repeated B times in order to obtain our vector of B permuted test statistic. In other words, we will obtain our vector $T = T_0, T^*_1, \cdots, T^*_b, \cdots, T^*_B$ of $B+1$ test statistic. Finally we applied our empirical cumulative function:

$$\hat{L}(t) = \frac{\frac{1}{2} + \frac{1}{2}\sum_{b=1}^{B} \mathbf{I}\left(T^*_b = t\right) + \sum_{b=1}^{B} \mathbf{I}\left(T^*_b < t\right)}{B+1}, b = 1, \cdots, B \tag{5}$$

to retrieve the p-value $\lambda = \hat{L}(T^0)$ and its simulated distribution $\lambda^*_1, \cdots, \lambda^*_b, \cdots, \lambda^*_B$.

As said in Eq. 4, we need to perform $S-1$ two sample problems and combine their results to obtain our global p-value λ'. Be $\lambda_1, \cdots, \lambda_m, \cdots, \lambda_{S-1}$ the p-values obtained from the $S-1$ two sample problems and be $\lambda^*_{m,1}, \cdots, \lambda^*_{m,B}$ the simulated distribution of the m-th p-value. A combining function θ is then

applied to obtain $T_0' = \theta(\lambda_1, \cdots, \lambda_m, \cdots, \lambda_{S-1})$ and its simulated distribution $T_b' = \theta(\lambda_{1,b}^*, \cdots, \lambda_{m,b}^*, \cdots, \lambda_{S-1,b}^*)$.

We then computed our global p-value $\lambda' = \hat{L}'(T_0')$, where

$$\hat{L}'(t) = \frac{\frac{1}{2} + \frac{1}{2}\sum_{b=1}^{B} \mathbf{I}\left(T_b' = t\right) + \sum_{b=1}^{B} \mathbf{I}\left(T_b' < t\right)}{B+1}, b = 1, \cdots, B \tag{6}$$

3 Case Study

The research and development team of a company aims to investigate whether the design of a new website can have a positive impact on users' overall satisfaction. To achieve this goal, among others, they conduct a survey involving participants who navigate the same website with different designs. For confidentiality purposes related to our case study, we will indicate three generic versions for the website, namely Design A, Design B, and Design C. Users' overall satisfaction is measured using a 10-point slider scale. We applied the four procedures proposed by the competitors and the NPC based approach examined in the previous Sect. 2.

3.1 Results and Discussion

Design A represents the old website design, Design B exhibits some layout changes, while Design C is entirely new. The idea is that the average users' overall satisfaction for Design A will be lower compared to Design B and Design C. This is confirmed by the descriptive statistics in Fig. 1.

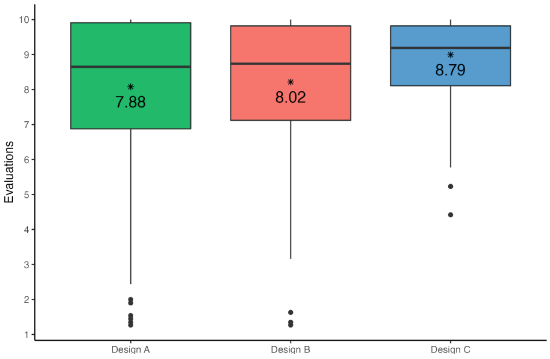


Fig. 1. Boxplots representing customers' evaluations on a 10-point slider scale, for the three distinct design options.

Hence, the objective is to verify whether these three groups exhibit statistically significant differences based on the overall ratings they have received. In

other words, we are interested in testing the null hypothesis $H_0 : \mu_A = \mu_B = \mu_C$, against the alternative $H_1 : \mu_A \leq \mu_B \leq \mu_C$. The descriptive statistic shows that the group variances are unequal, confirming that this is a common occurrence in practice. This is supported by the results of Levene's Test for Homogeneity of Variance, yielding a p-value of 0.0032.

Focusing on the results related to our system of hypothesis, the NPC methodology returns a p-value of 0.0014 when using the mean as the test statistic, and a p-value of 0.002 when employing the Welch t-test as the test statistic. In both cases, considering an α-level of 0.05, there is sufficient statistical evidence to reject the null hypothesis. Therefore, we can conclude that there is a statistically significant difference between the average users' overall satisfaction in the three considered groups, in particular the average users' overall satisfaction for Design A is lower compared to Design B and Design C.

Regarding the competitors, both the test statistic and critical values are obtained by setting $\alpha = 0.05$, as detailed in the following Table:

Table 1. The table presents the results in terms of the test statistic values and critical values for the 4 competitor approaches.

Method	Test statistic	Critical value
MinT	0.4049	0.3537
AMinT	0.4049	0.3623
LRT	0.0010	0.1502
ALRT	13.80	3.841

As shown in Table 1, for MinT, AMinT, and ALRT, the test statistic exceeds the critical value, indicating sufficient statistical evidence to reject the null hypothesis. Similarly, for LRT, the test statistic is lower than its critical value, leading to the same conclusion.

We can conclude that both the parametric bootstrap approaches and the permutation-based approach yield similar conclusions. However, the permutation-based approach shows the advantage of providing a p-value, which offers more informative results compared to solely the test statistic and its critical value. A simulation study is currently being conducted demonstrating very good properties of the NPC methodology, both in case of power and size performances.

4 Conclusion

In conclusion, the aim of this short paper is to focus on testing the null hypothesis of equals means against ordered alternatives in a one-way design with unequal variances. While recently Mondal et al. (2023) [6] propose a parametric bootstrap approach, we introduce an alternative using the permutation approach based on

the NPC methodology. Indeed, the NPC methodology presents promising advantages. It is a distribution-free technique, and assuming that the exchangeability assumption is justified under the null, along with the non-intersection of cumulative distribution functions under the alternative hypothesis, it offers an exact solution. We compare the two approaches in a case study about the centered around the introduction of a new service, particularly examining the impact of the design of a website on the overall customer experience. Both the parametric bootstrap procedures and the NPC-based methodology yield the same conclusions. A simulation study is currently underway.

Acknowledgements. The authors carried out this studies within the MICS (Made in Italy - Circular and Sustainable) Extended Partnership and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 - D.D. 1551.11-10-2022, PE00000004), the MOST - Sustainable Mobility National Research Centre and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 - D.D. 1033.17-06-2022, CN00000023) and BIRD 2022 project titled'Fuzzy theory in Unsupervised Machine Learning algorithm and Sentiment Analysis. This manuscript reflects only the author's views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

1. Bishop, T.A., Dudewicz, E.J.: Exact analysis of variance with unequal variances: test procedures and tables. Technometrics **20**(4), 419–430 (1978)
2. Bishop, T.A., Dudewicz, E.J.: Heteroscedastic anova. Sankhyā: Indian J. Stat. Ser. B 40–57 (1981)
3. Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R. John Wiley & Sons (2014)
4. Chen, S.Y.: One-sided range test for testing against an ordered alternative under heteroscedasticity. Commun. Stat.-Simul. Comput. **29**(4), 1255–1272 (2000)
5. Chen, S.Y.: One-stage and two-stage statistical inference under heteroscedasticity. Commun. Stat.-Simul. Comput. **30**(4), 991–1009 (2001)
6. Mondal, A., Pauly, M., Kumar, S.: Testing for ordered alternatives in heteroscedastic ANOVA under normality. Stat. Pap. **64**(6), 1913–1941 (2023)
7. Noguchi, K., Konietschke, F., Marmolejo-Ramos, F., Pauly, M.: Permutation tests are robust and powerful at 0.5% and 5% significance levels. Behav. Res. Methods **53**(6), 2712–2724 (2021). https://doi.org/10.3758/s13428-021-01595-5
8. Oshima, T., Algina, J.: Type I error rates for James's second-order test and Wilcox's hm test under heteroscedasticity and non-normality. Br. J. Math. Stat. Psychol. **45**(2), 255–263 (1992)
9. Pauly, M., Asendorf, T., Konietschke, F.: Permutation-based inference for the AUC: a unified approach for continuous and discontinuous data. Biom. J. **58**(6), 1319–1337 (2016)
10. Pauly, M., Brunner, E., Konietschke, F.: Asymptotic permutation tests in general factorial designs. J. R. Stat. Soc. Ser. B Stat Methodol. **77**(2), 461–473 (2015)

11. Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. John Wiley & Sons (2010)
12. Wilcox, R.R.: New designs in analysis of variance. Annu. Rev. Psychol. **38**(1), 29–60 (1987)
13. Wilcox, R.R.: ANOVA: the practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. Br. J. Math. Stat. Psychol. **48**(1), 99–114 (1995)

Evaluating the Geographical Patterns of Intimate Partner Violence in Nigeria, Zambia, and Zimbabwe Using INLA-SPDE Approach

Micaela Arcaio[1](✉), Anna Maria Parroco[1], and Chibuzor Christopher Nnanatu[2]

[1] University of Palermo, Palermo, Italy
micaela.arcaio@unipa.it
[2] WorldPop Group, University of Southampton, Southampton, UK

Abstract. Recent data show that one in three women in Sub-Saharan Africa is abused at least once in her lifetime by an intimate partner. The fight against violence against women has thus been included among the targets of two Sustainable Development Goals of the United Nations.

The literature shows that there exists a complex interplay between personal and contextual characteristics, and where community-level characteristics such as living in rural areas come into play, there is a higher presence of intimate partner violence.

The aim of this preliminary study is to examine the geographical patterns of the number of victims of intimate partner violence in Nigeria, Zambia, and Zimbabwe, using the Demographic and Health Survey (DHS) data. Indeed, this survey provides data on current and past experiences of abuse and the geolocation of its clusters. These data were used in a Bayesian hierarchical modelling approach within the framework of INLA-SPDE, to model the total number of victims at the cluster level – which is assumed to be continuously observed in an underlying Gaussian random field and with a Poisson distribution. The models were explored with one final model selected for each country, and posterior estimates of the model parameter values (along with the 95% credible intervals) were analysed.

Preliminary results show that people living in rural areas are more exposed to violence. They also provide a useful evidence base for policymakers to create programs centred around the people living in the identified hotspots whilst taking specific cultural practices of their residents into consideration.

Keywords: Intimate partner violence · INLA-SPDE · Sub-Saharan Africa

1 Introduction

Intimate partner violence (IPV) is present in every Country in the world [1]. IPV, or domestic violence, has been defined as any act of physical, psychological, or sexual abuse by an intimate partner, resulting from patterns of control and power [2]. Although both men and women can be affected by this form of violence, most victims are women, with men as the main perpetrators [3, 4]. It is estimated to be more widely spread in the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 53–57, 2025.
https://doi.org/10.1007/978-3-031-64431-3_9

least developed regions of the world. In Sub-Saharan Africa, for example, the prevalence of lifetime IPV – i.e. at least once in the lifetime of a woman aged 15–49 years – is 33% [1].

The United Nations has included the fight against violence against women in two of the Sustainable Development Goals (SDG), both in the "Gender Equality" goal and in "Peace, Justice and Strong Institutions", calling for the end of this phenomenon [5].

The literature on IPV often reiterates how relevant the interplay between personal and contextual characteristics is [6, 7]. For this specific reason, an ecological model is used to describe the way IPV's drivers act at different levels of influence, and how they affect one another as well [8]. Among others, community-level characteristics, such as poverty and living in rural areas, have been found to facilitate patterns of violence [9, 10].

The aim of this preliminary study is to investigate the geographical patterns of the presence of victims of IPV in Nigeria, Zambia, and Zimbabwe, using data from the Demographic and Health Survey. The choice of these three countries is due to data availability about current experiences of abuse on female respondents by the partner and the geolocalisation of clusters of households of residence in the years 2013–2023 among the first twenty countries for HIV prevalence.

2 Data and Methods

The data used in this analysis are drawn from the Demographic and Health Survey (DHS) Program. The DHS is a nationally representative household survey, that employs harmonized questionnaires to improve the collection, analysis and dissemination of demographic and health data. Eligible respondents for individual interviews are identified through the sampled households. In this study, we use the results of the women's questionnaire, in which the responses to a special module to investigate past and current experiences of abuse perpetrated by intimate partners, family members, and other individuals are contained. In this case, we focused on currently married women and on instances of abuse that happened in the twelve months before the interview. The GPS coordinates of the clusters of households, also available in the DHS, were used in the analysis in the spatial component of the model.

Here, all the latest DHS surveys conducted in Nigeria, Zambia, and Zimbabwe in 2013–2023 are involved.

Given the aim of this analysis, we implement a Bayesian hierarchical regression model within the Integrated nested Laplace approximation in conjunction with the Stochastic Partial Differential Equation (INLA-SPDE) technique [11, 12]. This is used to account for the geographical characteristics of the presence of victims of IPV in the countries of interest.

The response variable is the count of the number of women who have been abused at least once in the past year (either physically, emotionally, or sexually, or any combination of the three) at the cluster level for each dataset. We assume that the number of people who were abused is spatially continuous and it can be modelled using a Gaussian random field.

The fitted spatial model will use the stochastic partial differential equation approach (SPDE) using the R-INLA package [13], so that we can then predict how many women

are abused at unsampled locations. The SPDE approach approximates the Gaussian field as a discrete Gaussian Markov random field via a finite basis function defined on a triangulated mesh of each country; the meshes are built here based on a non-convex boundary for the coordinates and the smoothness parameter to build the SPDE model on the mesh is set equal to 1.

We assume that the response follows a Poisson distribution and its response function is given by the sum of an intercept, and a random effect given by the region or state of residence (data on state of residence was only available for Nigeria, thus this kind of sub-regional level of analysis was employed for this country), and a spatially structured random effect that follows a zero-mean Gaussian process with Matérn covariance function.

Table 1 shows some descriptive statistics of the samples examined.

Table 1. Descriptive statistics of the datasets

Variable	Nigeria	Zambia	Zimbabwe
Year of interview	2018	2018	2015
Average number of abused women in the clusters	6.54	12.15	12.93
SD of number of abused women in the clusters	5.05	9.36	8.19

3 Results

Here we present the posteriors of the predictions. In particular, we use the posterior mean and the lower and upper limits of the 95% credible intervals. In this case, the lower limit represents the "best possible scenario under the model", i.e. the lowest predicted number of abused women for that specific point in the map, whereas the upper limit is the "worst possible scenario under the model", the highest possible number of victims for that model. These values are then represented in the maps below to investigate the presence of hotspots (the maps on the left represent the lower limits, the ones in the centre represent the predicted means and the ones on the right are the upper limits).

From the maps, some patterns are visible for all countries. Noticeably, for Nigeria, there seem to be several hotspots in the North-Western region and the Borno state; in Zambia, hotspots seem to be concentrated in the South-Eastern border of the country, while in Zimbabwe, multiple hotspots can be found all over the country (Figs. 1, 2 and 3).

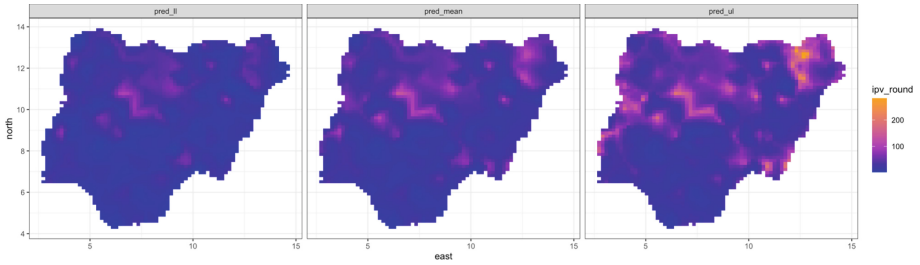


Fig. 1. Possible scenarios for hotspots of the number of victims of IPV in Nigeria in 2018

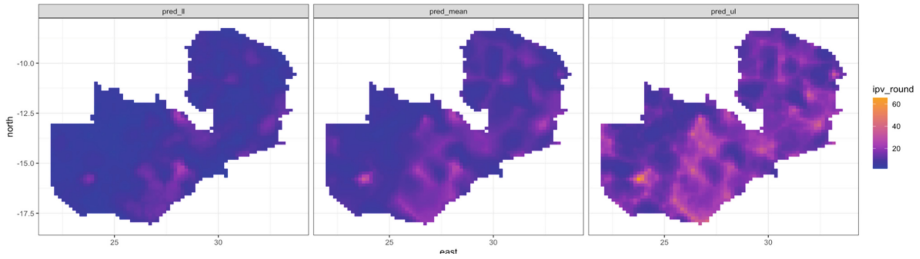


Fig. 2. Possible scenarios for hotspots of the number of victims of IPV in Zambia in 2018

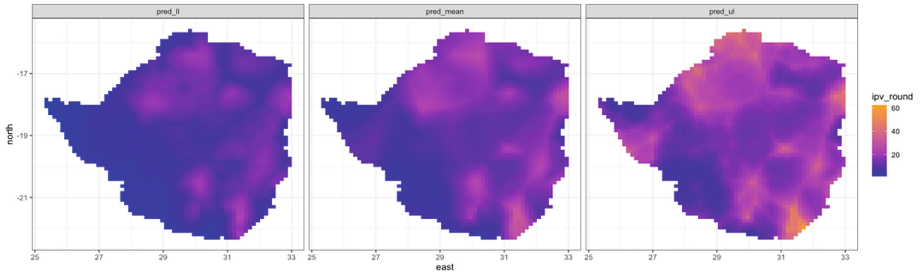


Fig. 3. Possible scenarios for hotspots of the number of victims of IPV in Zimbabwe 2015

4 Conclusions

While the entire world is affected by intimate partner violence, it is particularly relevant in developing countries. To improve women's conditions overall, indeed, the United Nations have included the fight against violence against women among their targets of the Sustainable Development Goals.

Bayesian hierarchical models (with the INLA-SPDE approach, since the data are geostatistical) give the opportunity to check for geographical patterns within countries – in this case, to check for hotspots with a high number of abused women. This approach offers a better solution than simple descriptive maps and a robust alternative to other models because it allows to take spatial correlation into account, as well as a formal expression of uncertainty in the estimates.

In this preliminary study, we modelled the number of victims of IPV at the cluster level in Nigeria, Zambia and Zimbabwe using data from the DHS. The posteriors of the predictions of the estimated models allowed to find some hotspots in the countries. Noticeably, what really jumped from the results, is the mapping of Nigeria, where desertic regions in the North seem to be particularly characterised by this phenomenon.

This work is not exempt from limitations. Overdispersion may affect the data, thus changing the response distribution may help model the number of victims better. Moreover, these models are currently being explored to control for some characteristics of respondents and other relevant characteristics of the clusters.

Policy-makers could, however, find this type of methodology particularly useful, since it would allow to implement programs centred around the people to be involved, taking specific cultural practices of their residents into consideration. Building policies around the people living in particularly affected areas would also help reduce inequalities that may go undetected when studying phenomena at a national level.

References

1. World Health Organization: Violence against women prevalence estimates, 2018: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women. https://www.who.int/publications/i/item/9789240022256. Accessed 15 Feb 2024
2. United Nations: What is domestic abuse?. https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Accessed 15 Feb 2024
3. Centers for Disease Control and Prevention: Fast facts: Preventing intimate partner violence. https://www.cdc.gov/violenceprevention/intimatepartnerviolence. Accessed 15 Feb 2024
4. World Health Organisation: Understanding and addressing violence against women: Intimate partner violence. https://apps.who.int/iris/handle/10665/77432. Accessed 15 Feb 2024
5. United Nations, Take Action for the Sustainable Development Goals. https://www.un.org/sustainabledev.Aelopment/sustainable-development-goals/. Accessed 15 Feb 2024
6. Oyediran, K.A., Feyisetan, B.: Prevalence and contextual determinants of intimate partner violence in Nigeria. Afr. Popul. Stud. **31**(1) (2017)
7. Gage, A.J.: Women's experience of intimate partner violence in Haiti. Soc Sci Med **61**(2), 343–364 (2005)
8. Centers for Disease Control and Prevention: The social-ecological model: A framework for prevention. https://www.cdc.gov/violenceprevention/about/social-ecologicalmodel.html. Accessed 15 Feb 2024
9. Jeyaseelan, L., Kumar, S., Nelakantana, N., Peedicavil, A., Pillai, R., Duyyury, N.: Physical spousal violence against women in India: some risk factors. J. Biosoc. Sci. **39**(5), 657–670 (2007)
10. Deyessa, N., Berhne, Y., Ellsberg, M., Emmelin, M., Kullgren, G., Högberg, U.: Violence against women in relation to literacy and area of residence in Ethiopia. Glob. Health Action **3**(1) (2010)
11. Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Series B (Stat. Methodol.) **73**(4), 423–498 (2011)
12. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Series B (Stat. Methodol.) **21**(2), 319–392 (2009)
13. R-INLA. https://www.r-inla.org/home. Accessed 15 Feb 2024



A Two-Part Beta Regression Model with Measurement Error

Maria Felice Arezzo(✉), Giuseppina Guagnano, and Domenico Vitale

Sapienza University of Rome, Dept. MEMOTEF, Rome, Italy
mariafelice.arezzo@uniroma1.it

Abstract. There are phenomena bounded into positive intervals which show an excessive number of zero values, making it difficult the application of standard modeling procedures. To overcome such a problem, a two-part model is often employed. The model involves two stochastic models: the first governs whether the response variable is zero or positive and the second, conditional on its being positive, models the bounded variable. In this work, we extend this modeling framework to cope with measurement errors affecting the dependent variables. The performance of the proposed approach compared to a *naive* procedure, which ignores the measurement errors, is evaluated through a Monte Carlo simulation study.

Keywords: Two-part model · beta regression · measurement error

1 Introduction

In many applied work and especially in economics, it is very common to come across data that is measured as proportions with a high mass at 0: the percentage of women in the public sector workforce, the proportion of income spent on tobacco or on medical expenses, the percentage of hours worked paid in cash, just to name a few. In all these situations, the variable of interest Y is bounded in the $[0, 1)$ interval. The linear regression model is not appropriate, as it may produce fitted values outside such interval. A possible solution is to transform the dependent variable so that it assumes values on the real line, and then model the mean of the transformed response as a linear predictor based on a set of exogenous variables. Unfortunately, this approach has drawbacks: the model parameters lack interpretations in terms of the original response and, since the measures of proportions typically display asymmetry, inference based on the normality assumption can be misleading.

When data of this type has a high mass at zero, it adds another layer of complexity because estimating the regression coefficients results in a poor fit. A popular solution to model non-negative and bounded continuous variable with a mass at 0 is the two-part model. In such framework, the first part models whether the response variable is zero or positive and the second part explains the level, conditional on its being positive.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 58–63, 2025.
https://doi.org/10.1007/978-3-031-64431-3_10

In the setting of interest, a two-part Beta model has the following form:

$$Y = 0 \quad \text{with probability} \quad 1 - \pi \tag{1}$$

$$\sim Beta\,(\mu\phi; (1-\mu)\phi) \quad \text{with probability} \quad \pi \tag{2}$$

where μ is the mean of the response variable and ϕ can be interpreted as a precision parameter in the sense that, for fixed μ, the larger the value of ϕ, the smaller the variance of Y. For the generic $i-th$ observation, both the probability π_i and the expected value μ_i can be explained through a set of covariates in a regression setting. The resulting model is the so called two-part Beta regression:

$$\text{Part one:} \qquad P(Y_i = 0|\boldsymbol{Z}_i) = 1 - P(Y_i = 1|\boldsymbol{Z}_i) = 1 - \pi_i = 1 - \Phi(\boldsymbol{Z}_i'\boldsymbol{\beta}) \tag{3}$$

$$\text{Part two:} \qquad E(Y_i|Y_i > 0, \boldsymbol{X}_i) = \mu_i = \frac{exp\,(\boldsymbol{X}_i'\boldsymbol{\theta})}{1 + exp\,(\boldsymbol{X}_i'\boldsymbol{\theta})} \tag{4}$$

where $\Phi(.)$ is the standard normal cdf. In applied works, it is common for data to be flawed with measurement error. This can occur during an interview if the respondent misunderstands the question or deliberately provides false information. The two-part Beta regression model has been extended to account for two types of measurement error that affect the response variable. The first type determines whether it is zero or positive, while the second its level.

Let us introduce some notation on the measurement error models used in our work. Let Y^O be the fallible/error-prone variable and y^O be the observed value. The misclassification model, which specifies the behaviour of Y^O given the true unobserved value $Y^T = y^T$, is characterized by the probability:

$$P(Y^O = y^O|Y^T = y^T), \tag{5}$$

which defines misclassification probabilities when $y^T \neq y^O$ [5].

In many real situations, it is plausible to assume the occurrence of only one type of misclassification [2]. For example, in a survey where respondents are asked about their involvement in deviant behaviours, they are likely to conceal such behaviours due to the social desirability bias [1]. As such, in our work we consider only $P(Y^O = 0|Y^T > 0) = \alpha$. Considering the regression setting in (3), it follows that:

$$P(Y_i^O > 0|\boldsymbol{Z}_i) = (1-\alpha)\pi_i = (1-\alpha)\Phi\,(\boldsymbol{Z}_i'\boldsymbol{\beta})\,. \tag{6}$$

We also admit a measurement error on the continuous part, e_i, which may depend on individual characteristics:

$$e_i = \frac{exp\,(\boldsymbol{W}_i'\boldsymbol{\gamma})}{1 + exp\,(\boldsymbol{W}_i'\boldsymbol{\gamma})}. \tag{7}$$

The conditional expected value of Y_i^O then becomes:

$$\mu_i^* = \frac{exp\,(\boldsymbol{X}_i'\boldsymbol{\theta})}{1 + exp\,(\boldsymbol{X}_i'\boldsymbol{\theta})} \cdot e_i. \tag{8}$$

2 The Proposed Model

The two-part model for the error free dependent variable Y^T, recalled in (3)–(4), has the following mixture p.d.f. and likelihood [3]:

$$f\left(Y_i^T\right) = (1-\pi_i)I(Y_i^T = 0) + \pi_i g(Y_i^T | Y_i^T > 0, \boldsymbol{X}_i) \tag{9}$$

$$L(\beta, \theta) = \underbrace{\left[\prod_{Y^T=0}(1-\pi_i) \cdot \prod_{Y^T>0}\pi_i\right]}_{L_1(\boldsymbol{\beta})} \cdot \underbrace{\left[\prod_{Y^T>0} g(Y_i^T | Y_i^T > 0, \boldsymbol{X}_i)\right]}_{L_2(\boldsymbol{\theta})} \tag{10}$$

where β and θ are vectors of parameters that govern the part one and part two respectively, and $I(Y_i^T = 0)$ is an indicator function such that it equals 1 if $Y_i^T = 0$ and 0 otherwise; it is motivated by the fact that when $Y_i^T = 0$, its density collapses to a unit probability mass.

More precisely, $L_1(\beta)$ is the likelihood of a standard binary regression model and the corresponding link function is usually specified as probit or logit. $L_2(\theta)$ refers to the regression model for the continuous variable Y^T. In this work, we model L_1 with a probit link and L_2 as a Beta regression.

Since both parts are assumed to be affected by measurement errors, the homologous of the p.d.f. (9) for the error-prone Y^O is:

$$f\left(Y_i^O\right) = \sum_{j=1}^{4} \psi_{ji} \cdot g_j\left(Y_i^O\right), \tag{11}$$

where the weights ψ_{ji} are defined as:

$$\begin{aligned}
\psi_{1i} &= P(Y_i^O = 0, Y_i^T = 0) = (1-\pi_i), \\
\psi_{2i} &= P(Y_i^O = 1, Y_i^T = 1) = (1-\alpha) \cdot \pi_i, \\
\psi_{3i} &= P(Y_i^O = 1, Y_i^T = 0) = 0, \\
\psi_{4i} &= P(Y_i^O = 0, Y_i^T = 1) = \alpha \cdot \pi_i,
\end{aligned} \tag{12}$$

and the conditional densities $g_j\left(Y_i^O\right)$ are:

$$\begin{aligned}
g_1\left(Y_i^O\right) &= f\left(Y_i^O | Y_i^O = 0, Y_i^T = 0\right) = 1, \\
g_2\left(Y_i^O\right) &= f\left(Y_i^O | Y_i^O > 0, Y_i^T > 0, \boldsymbol{X}_i\right), \\
g_3\left(Y_i^O\right) &= f\left(Y_i^O | Y_i^O > 0, Y_i^T = 0\right), \\
g_4\left(Y_i^O\right) &= f\left(Y_i^O | Y_i^O = 0, Y_i^T > 0\right) = 1.
\end{aligned} \tag{13}$$

The density g_2 represents the main contribution in explaining Y_i^O, but its weight ψ_2 tends to zero as α goes to 1; g_3 is irrelevant because $\psi_3 = 0$.

The contribution of the $i-th$ unit to the likelihood is:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \alpha) = \{(1-\pi_i) + \alpha \cdot \pi_i\}^{(1-y_i^O)} \cdot \{(1-\alpha) \cdot \pi_i \cdot Beta\,(\mu_i^* \phi; (1-\mu_i^*)\phi)\}^{y_i^O} \tag{14}$$

where $Beta\,(\mu_i^* \phi; (1-\mu_i^*)\phi)$ is the density g_2 and μ_i^* is defined in Eq. (8).

3 Simulation Study

We present the finite sample performances of the proposed model and compare them to the two-part Beta regression model without measurement error, via Monte Carlo simulations. We assumed the following generating model for the error-free dependent variables:

$$\Pr\left(Y^T > 0 | Z_1, Z_2, Z_3\right) = \Phi\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\right), \tag{15}$$

$$E(Y^T | Y^T > 0) = \mu = \frac{exp(\theta_0 + \theta_1 X_1 + \theta_2 X_2)}{1 + exp(\theta_0 + \theta_1 X_1 + \theta_2 X_2)}. \tag{16}$$

The covariates are generated as follows: Z_1 is log-normal with zero mean and unit variance, Z_2 is binomial with $p = 1/3$, Z_3 and X_1 are uniformly distributed over the unit interval and X_2 has a standard normal distribution.

The observable variable, Y^O, is generated by setting to 0 a proportion α of cases with $y^T > 0$. We set $\alpha \in \{0.2, 0.4\}$. Finally, to generate the mismeasured continuous part we used equations (7) and (8), setting $\gamma \in \{0.5, 0.7\}$ and $W \sim Mult_4(p_1 = 0.01; p_2 = 0.06; p_3 = 0.33, p_4 = 0.60)$. The values of γ govern the under-reported proportion (UR) of the continuous part and the choice made for the simulations correspond approximately to UR $= (20\%, 10\%)$. For example, in a survey to determine the percentage of work paid outside the envelope, respondents would under-report this amount by an average of 20% if $\gamma = 0.5$. Across simulations we fixed: $\beta' = (-1, 0.2, 1.5, -0.6)$, $\theta' = (0.5, 1, 2)$, $\phi = 200$. We repeated each simulation scenario 200 times with samples of size $n = 5,000$.

The performance of the proposed approach, was assessed in terms of relative bias and compared with that of the traditional two-part Beta regression (naive). Results are summarized in Fig 1 for α and γ and in Fig. 2 for the regression parameters. The estimates of α and γ resulted in strict agreement with the true values. Referring to the regression parameters, for the first part of the model, ordinary probit produces estimates that are biased by more than 20%. As expected, the problem worsens as the amount of misclassification grows. Conversely, the proposed model provides unbiased estimates, for all levels of misclassification considered in this work. For the continuous part, the results of the proposed model are good since the estimates of all parameters are trustworthy, whereas those of the naive are severely biased (even more than 50%).

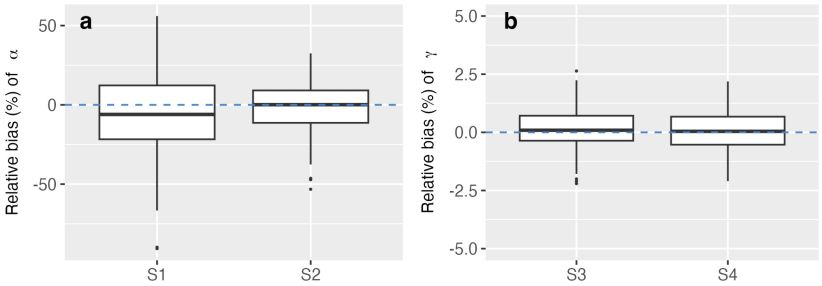


Fig. 1. Relative bias (%) for α (panel a) and γ (panel b) parameters estimated by the proposed approach over 200 simulation runs in different scenarios: S1: α=0.2; S2: α=0.4; S3: γ=0.7 (10% of UR rate); S4: γ=0.5 (20% of UR rate).

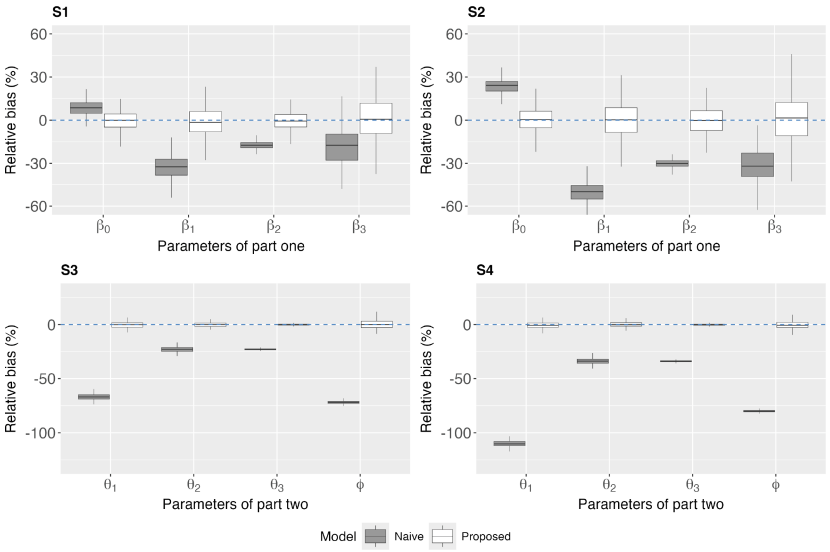


Fig. 2. Relative bias (%) of the parameter estimates in a two-part Beta regression model based on the proposed (white) and the naive (grey) approaches over 200 simulation runs in different scenarios: S1: α=0.2; S2: α=0.4; S3: γ=0.7 (10% of UR rate); S4: γ=0.5 (20% of UR rate).

4 Conclusions

In this work we propose an extension of the two-part Beta regression model allowing mismeasurement of the dependent variable both in the first and in the second part.

The main advantage of the proposed model is the reduction of the biases in the estimates that occur even when the error is of modest amount. This provides trustworthy estimates. Beyond that, there are other advantages in our model. In particular, we provide estimates of the misclassification probability α and of an

indirect evaluation (through γ) of the under reporting of the bounded continuous dependent variable. The quantification of under-reporting has important implications in many applied works such as those on deviant behaviors in economics. Relevant examples are in the context of tax evasion, undeclared and semi-formal work, envelope wages, drug consumption, bribery and so on. In such cases, the phenomena of interest are presumably measured with error, due to their sensitive and private nature. As such, we believe that our work contributes in filling an important gap in the applied statistical and economic literature.

References

1. Arezzo, M.F., Horodnic, I.A., Williams, C.C., Guagnano, G.: Measuring participation in undeclared work in Europe using survey data: a method for resolving social desirability bias. Socioecon. Plann. Sci. **91**, 101779 (2024)
2. Arezzo, M.F., Guagnano, G., Vitale, D.: Estimating the size of undeclared work from partially misclassified survey data via the EM algorithm. J. R. Stat. Soc. Ser. C Appl. Stat. **73**(3), 816–834(2024) forthcoming
3. Cragg, J.G.: Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica **39**, 829–844 (1971)
4. Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. J. Appl. Stat. **31**(7), 799–815 (2004)
5. Hausman, J.A., Abrevaya, J., Scott-Morton, F.M.: Misclassification of the dependent variable in a discrete-response setting. J. Econometrics **87**, 239–269 (1998)



TALL: A New Shiny App for Text Analysis

Massimo Aria[1], Corrado Cuccurullo[2], Luca D'Aniello[1], Michelangelo Misuraca[3], and Maria Spano[1(✉)]

[1] University of Naples Federico II, Naples, Italy
{massimo.aria,luca.daniello,maria.spano}@unina.it
[2] University of Campania Luigi Vanvitelli, Caserta, Italy
corrado.cuccurullo@unicampania.it
[3] University of Calabria, Arcavacata, Italy
michelangelo.misuraca@unical.it

Abstract. The rapid technological advancements allowed the processing of different kinds of data to study real-world phenomena in recent times. Within this context, textual data emerged as a crucial resource in several domains, paving the way for new research questions. Nonetheless, many researchers lack the programming skills needed to effectively analyse textual data, creating a demand for user-friendly analysis tools. Languages such as R and Python provide powerful and open-access solutions, but researchers often face time and resource constraints to become fairly proficient. This paper introduces TALL (*Text Analysis for All*), an R Shiny app encompassing a wide set of methods tailored for key text analysis tasks. Its aim is to address the needs of researchers, providing a versatile and general-purpose tool for text analysis and enabling them to extract valuable insights in a more viable and accessible manner.

Keywords: Natural Language Processing (NLP) · Text Analysis Tools · R Shiny App · User-friendly Interface

1 Introduction

In the big data era, researchers active in diverse domains face the challenge of analysing extensive datasets spanning textual data from survey responses, customer reviews, social media posts, and scientific articles. This new data source – more and more available for the increasing diffusion of digital platforms and online content – allows for valuable insights into several phenomena, ranging from social sciences to healthcare. Typically, researchers aim to identify patterns and trends in textual bodies, employing advanced *natural language processing* (NLP) approaches and machine learning algorithms to extract meaningful information and then explore new research questions, validate existing theories, and discover novel viewpoints. Computational tools facilitate the efficient processing and analysis of large text volumes, significantly reducing the time and effort required for qualitative methods. However, many researchers lack the necessary

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 64–70, 2025.
https://doi.org/10.1007/978-3-031-64431-3_11

programming skills for text analysis, creating a demand for user-friendly analytical tools. Despite the powerful capabilities of R and Python, acquiring proficiency in these programming languages often requires additional effort.

This paper introduces the first release of TALL (*Text Analysis for ALL*)[1] a new R Shiny app that incorporates all the primary text analysis approaches and their advances. TALL provides a comprehensive workflow for data cleaning, pre-processing, statistical analysis, and visualisation of textual data, combining state-of-the-art techniques and new proposals into an intuitive dashboard that enables the exploration of textual data and the execution of analyses without extensive programming knowledge.

2 Related Background

The interest in textual data is not new, dating back to the 1980 s. The development of tools capable of analysing the continuously growing amount of unstructured data coming from text written in natural language has been fundamental alongside the methodological progress in this field. Nowadays, R and Python offer the most trending tools for text analysis and visualisation. However, several analytical solutions for no-coders have been proposed over the years.

The software Spad (*Systéme portable d'analyse de données*) [7] and its extension devoted to textual data Spad-T [8] could be considered one of the first proposals for the statistical analysis of textual bodies in the middle 80 s. This resource was originally distributed in a freeware context and then commercially distributed. Spad-T offers a set of textual data analyses (in the wake of the French Data Analysis school) and simple tools for data manipulation. The most classic approach is Correspondence Analysis on lexical tables crossing groups of text chunks and their composing words. It is necessary to incorporate metadata to enable the possibility of analysing these data structures (i.e., additional information on the text chunks). This latter was one of the limitations of Spad-T, which was mainly developed to analyse open-ended questions, making the analysis of other types of text a bit cumbersome. Subsequent to factorial analysis, a Hierarchical Clustering [20] may be carried out on factor coordinates using the logic of the so-called tandem analysis. Analysing the per-cluster vocabulary and the most characteristic text chunks is also possible to aid interpretation.

In France, around the same time, was also developed Alceste (*Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte*) [17], with similar capabilities on the analytical side (based on factorial analyses and hierarchical clustering [16]), but more powerful regarding the pre-process of textual bodies and, even more critical, designed for all kinds of text. As in the case of Spad, this software was developed in a freeware context but later became a proprietary one. An open-access alternative – developed in R and Python – has been offered by IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*) [15]. The software provides some classical text statistics, correspondence analysis and clustering on textual bodies and segmented textual

[1] Available at: https://github.com/massimoaria/tall.

bodies, together with network tools useful for analysing *words* × *words* tables with community detection [3]. The documentation is comprehensive, but it can be overwhelming for beginners. Even if IRaMuTeQ is a powerful tool, it can be difficult to use due to the importing procedure requiring a specific text structure. On the other hand, it is possible to use the software on the main operating systems, allowing a wider diffusion in comparison with the other alternatives.

Similar to IRaMuTeQ, another open-access R-based software to be mentioned is KH Coder [5]. Differently from the other alternatives, developed in accordance with the principles of the French Data Analysis school, KH Coder offers, in addition to the above-mentioned approaches, the possibility of using multidimensional scaling to explore the association structure of data and more network analysis tools. Moreover, it allows performing linguistics analyses using some interesting resources such as the *Stanford POS Tagger*[2] and the *FreeLing* toolkit[3].

3 A Comprehensive Workflow for Text Analysis

TALL combines the functionality of a set of R packages developed for NLP tasks with the web app's ease of usage, thanks to the Shiny environment. The workflow aims to facilitate the exploration and analysis of textual data by systematically processing and inspecting the content of the analysed textual body.

The first step, $< Import\ and\ Edit >$, involves acquiring one or more text files in various formats (such as txt, csv, xlsx, and pdf) to easily load a wide range of textual data. Subsequently, text can be subjected to several editing actions, including splitting it into smaller segments, such as chapters or paragraphs, or selecting text subsets for sampling purposes. Users can supplement the imported text with additional information (e.g., author, publication date, rating) initially attached to the content or added during the analysis. Considering both the analysis goals and the availability of external variables, text can also be filtered, enabling the focus on specific groups for comparative purposes.

The $< Pre - processing >$ step allows the preparation of the text for the subsequent analysis. TALL was designed as a general-purpose software capable of analysing any type of text, from books to newspaper articles, blogs, scientific papers, reviews, and data from social media. Depending on the nature of the text, a cleaning step may be essential. In the text normalisation tab, users can decide which elements to remove from the text, as in the case of social media posts (URLs, Hashtags, YouTube IDs, Emoji, Tags, IP Addresses) or of other kinds of text (e-mails, extra spaces, non-ASCII characters, numbers, time, dates, zip postal codes, page abbreviations, bibliographic citations). Before beginning the annotation process (i.e., tokenisation, PoS tagging, and lemmatisation), a language model must be downloaded. TALL employs pre-trained models provided

[2] https://nlp.stanford.edu/software/tagger.shtml.
[3] https://nlp.lsi.upc.edu/freeling.

by the Universal Dependencies Treebanks (UDT)[4], a framework for grammatical annotation on different human languages. By using the models listed in the UDT, TALL is able to support the analysis of text written in 60 different languages. Each text is parsed into individual tokens (words) and tagged with its respective Part-of-Speech (PoS) label to better understand word usage patterns. All the statistical analyses could be performed alternatively on tokens or lemmas. Moreover, users can define and load custom lists of words for various research purposes (e.g., to substitute synonyms, remove domain stop-words, and semantically tag specialised terms). Another crucial aspect when we deal with text analysis is identifying and handling multi-word expressions and collocations. To face this issue, TALL implement the Rapid Automatic Keyword Extraction (RAKE) algorithm [18] that identifies candidate keywords and scores them using word co-occurrences. At the end of the pre-process, users can select specific PoS tags to focus their analyses considering only a set of them (e.g., nouns, adjectives, verbs, collocations, custom words).

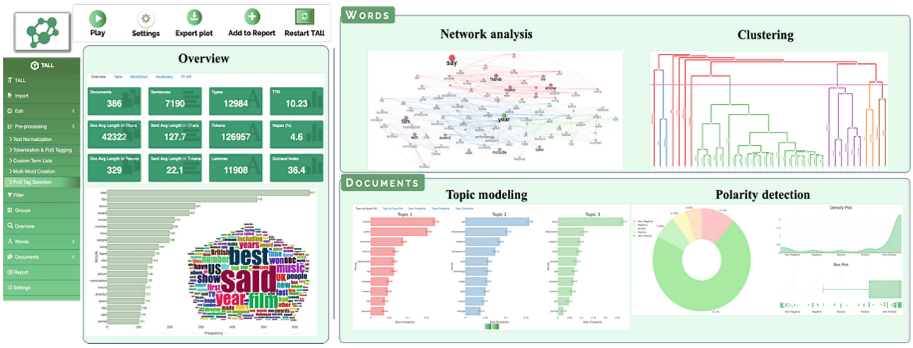


Fig. 1. Synopsis of the TALL analytics.

The < *Statistical analysis and Visualisation* > step opens the opportunity to explore the text by performing several approaches, as listed in Fig. 1. Descriptive statistics, word frequency distribution, and word-clouds provide an initial overview of the textual body. The rationale of the TALL analytical workflow is that it is possible to consider a word or document level. Analyses on the word-level include a set of statistical methods mainly devoted to topic detection. The most intuitive approach is to identify and visualise through dynamic plots the most frequent words for each PoS tag, looking at their absolute frequency but also considering more complex weighting schemes such as term frequency/inverse document frequency (TF-IDF) [19] to uncover words with a highest discriminating power. Despite its simplicity, looking at the word frequency distribution often gives a broad idea of the textual body content, but it is insufficient to identify topics. A topic can be represented as a set of meaningful words with syntagmatic

[4] https://universaldependencies.org.

relatedness. Following this definition, the three methods most widely used in the literature [12] are implemented in TALL:

- *Cluster Analysis*, to group words based on their usage patterns or context;
- *Correspondence Analysis*, to explore semantic relationships among words, identifying the latent structure of the textual body (often in combination with hierarchical clustering);
- *Network Analysis*, to analyse co-occurrence patterns of words within the text, highlighting subsets of words strictly related through community detection.

The document-level analyses include a set of statistical methods to cope with specific tasks where the focus is on the entire textual body:

- *Topic Modelling*, to identify both prominent topics and their distribution within documents using the well-known Latent Dirichlet Allocation [1]. TALL also allows to automatically determine the number of topics through some alternative metrics [2,4], but users can also explore different solutions setting by their own a given number of topics;
- *Polarity Detection*, to determine the positivity/negativity of documents choosing among various lexicons (e.g., Loughran and McDonald [9], NRC [13]);
- *Summarisation*, to summarise a text and rapidly capture key insights. TALL implements the TextRank algorithm [10], based on Google's PageRank [14], using sentence networks to extract the most relevant parts of the text.

The described workflow provides users with the key statistical methods to process text, facilitating and speeding up all the necessary steps and allowing them to share the results with others by downloading plots and reports.

4 Conclusions and Future Research

This paper briefly introduces the first version of TALL, a new Shiny app for importing, pre-processing, and analysing textual data. The idea of developing a new resource stems from the growing need of analysing textual content in several spheres, offering the opportunity to explore text quickly and efficiently, even for those who do not have programming skills. Using a user-friendly analytical tool, researchers can focus more on their domain expertise rather than spend significant time learning programming languages and writing complex code. A tool as TALL democratises text analysis, making it accessible to a broader audience and promoting interdisciplinary collaboration.

TALL is an ambitious project under continuous evolution. At the moment, all the implemented analyses rely on the Vector Space Model for the text algebraic representation. In this model, each document or text chunk is transformed into a p-dimensional vector, where p equals the number of unique words in the text (i.e., its vocabulary). Future developments will focus on implementing alternative approaches, such as *word embedding* [6]. This family of methods (e.g., Word2Vec

[11]) associates each word (or sequence of words) with a high-dimensional numerical vector reflecting their semantic relationship with other words, opening to the implementation of new statistical techniques.

Acknowledgement. This research has been financed by the following research projects: – PRIN 2022 *SCIK-HEALTH* (Project Code: 2022825Y5E - CUP: E53D23006110006); – PRIN 2022 PNRR *The value of scientific production for patient care in Academic Health Science Centres* (Project Code: P2022RF38Y - CUP: E53D23016650001).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S.: A density-based method for adaptive LDA model selection. Neurocomputing **72**, 1775–1781 (2009)
3. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. **101**, 5228–5235 (2004)
5. Higuchi, K.: New quantitative text analytical method and KH coder software. Jpn. Sociol. Rev. **68**, 334–350 (2017)
6. Hinton, G.E.: Learning distributed representations of concepts. In: Morris, R.G.M. (ed.) Parallel distributed processing: Implications for psychology and neurobiology, pp. 46–61. Clarendon Press/Oxford University Press (1989)
7. Lebartm, L., Morineau, A.: SPAD: Systéme Portable pour l'Analyse des Données, Technical report, CISIA (1985)
8. Lebart, L., Morineau, A., Bécue, M., M., Haeusler, L.: Introduction à SPAD-T intégré Version 1.5 P.C. Technical report, CISIA (1993)
9. Loughran, T., McDonald, B.: Textual analysis in accounting and finance: a survey. J. Account. Res. **54**, 1187–1230 (2016)
10. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411, Association for Computational Linguistics (2004)
11. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv (2013). https://doi.org/10.48550/arXiv.1309.4168
12. Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In: Iezzi, D.F., Mayaffre, D., Misuraca, M. (eds) Text Analytics, Advances and Challenges, pp. 17–28. Springer Cham (2020)
13. Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: using Mechanical Turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34. Association for Computational Linguistics (2010)
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project (1998)
15. Ratinaud, P.: IRAMUTEQ: Interface de R pour Les analyses multidimensionnelle de textes et de questionnaires http://www.iramuteq.org

16. Reinert, M.: Une methode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. Cah. Anal. Donn. **8**, 187–198 (1983)
17. Reinert, M.: Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. Bull. Sociol. Methodol. **26**, 24–54 (1990)
18. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: Berry, M.W., Kogan, J. (eds.) Text Mining: Applications and Theory, pp. 1–20. Wiley (2010)
19. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**, 513–523 (1988)
20. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Ann. Data Sci. **2**, 165–193 (2015)



On the Estimation of INAR Models with Pseudo-Variance Quasi-Maximum Likelihood

Mirko Armillotta[1,2(✉)] and Paolo Gorgi[1,2]

[1] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
m.armillotta@vu.nl
[2] Tinbergen Institute, Amsterdam, The Netherlands

Abstract. This work presents an application of the Pseudo-Variance Quasi-Maximum Likelihood Estimators (PVQMLE) for estimating Integer-valued Autoregressive (INAR) models. We apply this methodology to two diverse datasets: software download and animal skin lesions datasets. The approach is showed to outperform other quasi-likelihood methods, offering superior goodness-of-fit performance. Additionally, it enables the selection of thinning and innovation distributions within the INAR model framework. By proposing this new estimation technique, we enhance the accuracy and flexibility of modeling integer-valued time series data, thereby facilitating more robust analysis in various fields. This research contributes to advancing statistical methodologies for analyzing discrete-valued time series data, offering valuable insights into complex phenomena represented by such datasets.

Keywords: count time series · Gaussian quasi-likelihood · semi-parametric estimation

1 Integer-Valued Autoregressive Models

Consider a time series process $\{Y_t\}_{t\in\mathbb{Z}}$ of counts elements $Y_t \in \mathbb{N}$. One of the most popular approaches to model count time series is the Integer-valued Autoregressive (INAR) model [1]. The INAR(1) model is given by

$$Y_t = a \circ Y_{t-1} + \varepsilon_t\,, \quad t \in \mathbb{Z}, \tag{1}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an iid sequence of non-negative integer-valued random variables with mean $\omega_1 > 0$ and variance $\omega_2 > 0$, and '$\circ$' is the thinning operator proposed in [6]. For a given $N \in \mathbb{N}$, a general formulation of the thinning operator is $a \circ N = \sum_{j=1}^{N} X_j$ when $N > 0$, and 0 otherwise, where X_j is a sequence of iid non-negative integer-valued random variables following a distribution with finite mean a and variance b, say $X_j \sim D_X(a, b)$. A common formulation is the binomial thinning where X_j is a sequence of independent Bernoulli random variables with success probability $a \in (0, 1)$, therefore $a \circ N$ is a binomial random

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 71–76, 2025.
https://doi.org/10.1007/978-3-031-64431-3_12

variable with N trials and success probability a. The conditional mean of the INAR(1) is

$$\lambda_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1}) = aY_{t-1} + \omega_1 \,. \tag{2}$$

where $\mathcal{F}_{t-1} = \{Y_{t-1}, Y_{t-2}, \dots\}$ is the past information set. The form of the variance for INAR models is known to be linear in the observations, therefore the conditional variance can be specified with the form

$$\nu_t^* = bY_{t-1} + \omega_2. \tag{3}$$

However, the variance parameters $\gamma = (\omega_2, b)'$ are typically constrained to depend on the mean parameters $\psi = (\omega_1, a)'$ according to the chosen integer-valued distribution for the innovations ε_t and the thinning operator. For example, when $\varepsilon_t \sim Pois(\omega_1)$ and $\circ$ is the binomial thinning then $\omega_2 = \omega_1$ by the Poisson property and $b = a(1-a)$.

The model can be extended to general p-order. (INAR(p) process)

$$Y_t = a_1 \circ Y_{t-1} + \cdots + a_p \circ Y_{t-p} + \varepsilon_t \,, \quad t \in \mathbb{Z},$$

$$\lambda_t = \sum_{h=1}^{p} a_h Y_{t-h} + \omega_1 \,, \qquad \nu_t^* = \sum_{h=1}^{p} b_h Y_{t-h} + \omega_2.$$

The gold standard estimation of the unknown mean parameters ψ would be the conditional Maximum Likelihood Estimator (MLE) since it is asymptotically efficient. Nevertheless, the MLE in INAR models is well-known to be cumbersome and numerically difficult, especially when $p > 1$, due to nested summations of (p + 1)-fold convolution [5]. In such situations, the use of quasi-maximum likelihood methods becomes attractive.

2 Pseudo-Variance Quasi-Maximum Likelihood Estimators

The most used quasi-maximum likelihood method for INAR models is the conditional Least Squares Estimator (LSE). However, this is well-known to be inefficient since it employs only information about the first moment of the conditional distribution. Therefore, we propose to carry out estimation by using the novel Pseudo-Variance Quasi-Maximum Likelihood Estimators (PVQMLEs) recently introduced by [2]. For a sample of data $\{Y_t\}_{t\in[1,T]}$ these estimators appeal the Gaussian quasi-likelihood for ψ with a pseudo-variance scaling defined as

$$L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(-\log \nu_t^*(\gamma) - \frac{[Y_t - \lambda_t(\psi)]^2}{\nu_t^*(\gamma)}\right), \tag{4}$$

where λ_t is the true specification of the conditional mean whereas ν_t^* is called pseudo-variance as it is not necessarily the correctly specified conditional variance (ν_t). We denote the whole parameter vector that contains both the parameter of the mean and pseudo-variance with $\theta = (\psi', \gamma')'$ and $\theta \in \Theta = \Psi \times \Gamma \subset \mathbb{R}^m$

is its parameters space where m is the total number of parameters. Based on the quasi-likelihood function in (4), [2] define two estimators: Unrestricted and Restricted PVQMLE. The Unrestricted PVQMLE (U-PVQMLE) is based on the unconstrained maximization of the pseudo-likelihood without imposing any constrains between ψ and γ and is defined as

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L_T(\theta), \tag{5}$$

where $\hat{\theta} = (\hat{\psi}', \hat{\gamma}')'$ and $\hat{\psi}$ is the U-PVQMLE of ψ. The Restricted PVQMLE (R-PVQMLE) is the maximizer of (4) over a restricted parameter set Θ_R that imposes restrictions on the pseudo-variance parameters

$$\Theta_R = \{\theta \in \Theta : S\gamma = g(\psi)\},$$

where S is a selection matrix and g a continuous function. The estimator derived from the the set Θ_R is the R-PVQMLE,

$$\hat{\theta}_R = \text{argmax}_{\theta \in \Theta_R} L_T(\theta) \tag{6}$$

where $\hat{\theta}_R = (\hat{\psi}_R', \hat{\gamma}_R')'$ and $\hat{\psi}_R$ is the restricted PVQMLE of ψ.

Under suitable regularity conditions, [2] proved that the U-PVQMLE is at least as efficient as other quasi-likelihood methods proposed in the literature. Moreover, it may be expected that if the mean and pseudo-variance equations share common parameters in θ, like in INAR models, then ψ and γ are not completely distinct so that $\theta \in \Theta_R$, and the R-PVQMLE in (6) could show improved efficiency.

3 Results of INAR Models

In this paper, we apply PVQMLEs to the estimation of INAR models by employing (3) as the pseudo-variance for the U-PVQMLE. Moreover, we estimate a set of R-PVQMLEs where the parameters $\gamma = (\omega_2, b)'$ are subject to distinct restrictions according to different combinations of innovation distribution and thinning operator. In this way, we can identify the specification of INAR models that best fits the analyzed datasets. We present two real data applications showing the usefulness of the proposed methodology.

Application 1 (TEX downloads). The first application concerns the daily number of downloads of CBβ *TeXpert* TEX editor for Windows collected between June 1, 2006, and February 28, 2007 ($T = 267$). See Fig. 1. This dataset has been studied in several papers that employ INAR models with Bernoulli thinning, see [7] among others. For these data, we show that a different thinning like the Poisson or the geometric one leads to improved goodness of fit. The Poisson thinning implies $b = a$ whereas the geometric thinning leads to $b = a + a^2$ [4]. In this sense the former allows to introduce equidispersion ($\lambda_t = \nu_t^*$) wheres the latter induces overdispersion, i.e. $\lambda_t \leq \nu_t^*$. We also allow for different innovation distributions. Recall that when $\varepsilon_t \sim Pois(\omega_1)$ we have $\text{E}(\varepsilon_t) = \text{V}(\varepsilon_t) = \omega_1$ so the

model will impose the restriction $\omega_2 = \omega_1$. We can also consider the case where errors follow a negative binomial distribution $\varepsilon_t \sim NB(\omega_1, u)$ with $\mathrm{E}(\varepsilon_t) = \omega_1$ and $\mathrm{V}(\varepsilon_t) = \omega_1 + \omega_1^2/u$ so that also the innovations can be overdispersed ($\omega_2 > \omega_1$). In this case the restriction on the intercept will be $\omega_2 = \omega_1 + \omega_1^2/u$.

Table 1 reports the results of PVQMLEs for the parameter vector $\theta = (\omega_1, a, \omega_2, b)'$. The first row shows the results for the U-PVQMLE. The other rows present R-PVQMLE with different restrictions according to specific error distribution-thinning combination. The coefficients are all significant at 5% significance level. In the unrestricted model, $\hat{\omega}_1 < \hat{\omega}_2$ and $\hat{a} < \hat{b}$ suggesting that the binomial thinning formulation may be inadequate since in that case $b = a(1-a) < a$. Indeed, the BIC selects the restrictions induced by NB error distribution and geometric thinning as the best in goodness-of-fit terms, strictly followed by the unrestricted estimation. The AIC rather prefers the unrestricted model to the NB-Geom restricted one probably because the AIC tends to select models with more parameters. The performances of both models are really close and improved with respect to other restrictions. In general, the information criteria are lower for thinning and distributions generating a higher degree of dispersion. This provides an empirical evidence that the binomial thinning is less adequate than other thinning approaches for the analyzed data.

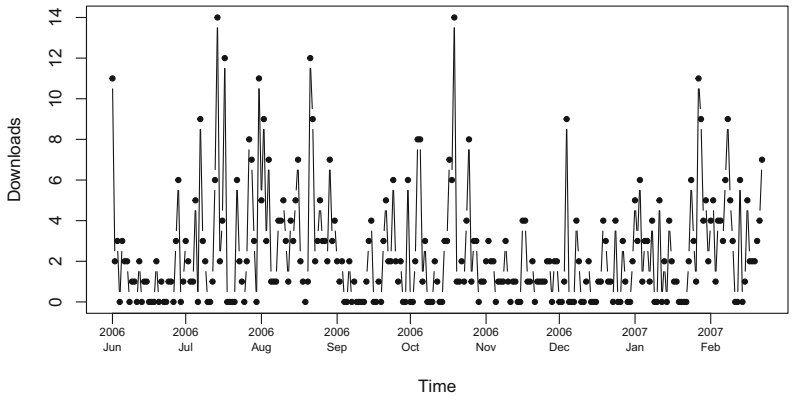


Fig. 1. Daily number of downloads of TEX editor 01/06/2006-28/02/2007.

Application 2 (Skin lesions). The second application studies the monthly number of submissions with skin lesions to animal health laboratories, 2003–2009, from a region in New Zealand (Fig. 2). The data set shows an inflation of zero meaning that the fraction of zero counts, f_0, in the time series is higher than the expected probability of "zero" which would occur in an INAR model where the errors ε_t were Poisson distributed. This latter probability equals $\exp(-\mathrm{E}(Y_t))$ and can be estimated by $\exp(-\bar{Y})$ where $\bar{Y}$ is the sample mean of Y_t. In our data set $f_0 = 0.405$ and $\exp(-\bar{Y}) = 0.240$ so an inflation of zeros is indeed present. The problem has been solved by several authors, starting from [3], by

Table 1. Estimation of TEX downloads data. Standard errors in brackets. Last two columns: Akaike and Bayesian information criteria. Lowest value in bold.

Model	Estimates					AIC	BIC
	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{a}$	$\hat{b}$	$\hat{u}$		
Unres.	1.673	4.586	0.292	0.933	-	**1258.27**	1272.62
	(0.215)	(1.186)	(0.081)	(0.463)			
Pois-Bin	2.552	-	0.205	-	-	1422.67	1429.85
	(0.260)		(0.058)				
NB-Bin	1.698	-	0.283	-	0.633	1264.56	1275.32
	(0.203)	-	(0.073)	-	(0.151)		
NB-Pois	1.641	-	0.311	-	0.658	1262.41	1273.18
	(0.231)	-	(0.092)	-	(0.160)		
NB-Geom	1.602	-	0.331	-	0.675	1260.42	**1271.78**
	(0.251)	-	(0.107)	-	(0.161)		

assuming a zero-inflated Poisson distribution of INAR errors, which induces over-dispersion on the error distribution, and a binomial thinning in (1). The estimation of the unrestricted model in Table 2 shows $\hat{\omega}_1 \approx \hat{\omega}_2$ and $\hat{a} < \hat{b}$. In fact, the model with the best performances according to information criteria is the one with Poisson errors and NB thinning. In other words, the overdispersion in the marginal distribution of the data seems to be generated by the thinning and not by the error distribution so the approach proposed in [3] does not find support by the empirical evidence of the data.

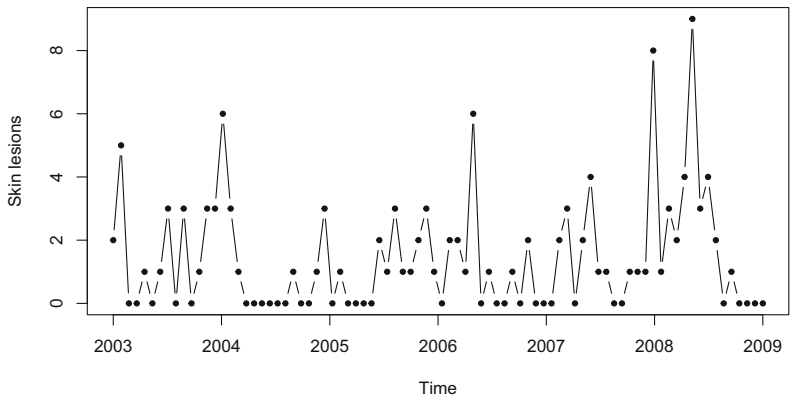


Fig. 2. Monthly number of submissions with skin lesions to animal health laboratories from a region in New Zealand during 2003–2009.

Table 2. Estimation of skin lesions data. Standard errors in brackets. Last two columns: Akaike and Bayesian information criteria. Lowest value in bold.

Model	Estimates					AIC	BIC
	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{a}$	$\hat{b}$	$\hat{u}$		
Unres.	0.771	0.747	0.450	2.223	-	315.78	325.51
	(0.154)	(0.196)	(0.148)	(0.917)			
Pois-Bin	1.400	-	0.273	-	-	352.80	357.66
	(0.291)		(0.099)				
NB-Bin	0.969	-	0.319	-	0.585	333.22	340.51
	(0.197)	-	(0.116)	-	(0.251)		
Pois-NB	0.733	-	0.627	-	-	**313.68**	**318.54**
	(0.144)	-	(0.115)	-	-		

This study introduced the estimation of INAR models through PVQMLEs to software downloads and animal skin lesions datasets. Restricted PVQMLEs showed the ability of enhancing model fit and allowing distribution selection flexibility. This improves accuracy in analyzing integer-valued time series data across various fields thereby advancing statistical methodologies for count time series. Future extensions can consider the development of efficient estimation tools for multivariate INAR models.

Acknowledgments. Mirko Armillotta acknowledges financial support from the EU Horizon Europe programme under the Marie Skłodowska-Curie grant agreement No.101108797.

References

1. Al-Osh, M., Alzaid, A.A.: First-order integer-valued autoregressive (INAR (1)) process. J. Time Ser. Anal. **8**, 261–275 (1987)
2. Armillotta, M., Gorgi, P.: Pseudo-variance quasi-maximum likelihood estimation of semi-parametric time series models. arXiv:2309.06100 (2023)
3. Jazi, M.A., Jones, G., Lai, C.-D.: First-order integer valued AR processes with zero inflated Poisson innovations. J. Time Ser. Anal. **33**(6), 954–963 (2012)
4. Lu, Y.: The predictive distributions of thinning-based count processes. Scand. J. Stat. **48**, 42–67 (2021)
5. Pedeli, X., Davison, A.C., Fokianos, K.: Likelihood estimation for the INAR(p) model by saddlepoint approximation. J. Am. Stat. Assoc. **110**(511), 1229–1238 (2015)
6. Steutel, F., Van Harn, K.: Discrete analogues of self-decomposability and stability. Ann. Probab. **7**(5), 893–899 (1979)
7. Weiß, C.H.: An Introduction to Discrete-valued Time Series. John Wiley & Sons (2018)

Estimating Spatial Anisotropy in Semiparametric Regression with Differential Regularization

Eleonora Arnone[1(✉)], Matteo Tomasetto[2], and Laura M Sangalli[3]

[1] Università di Torino, Turin, Italy
eleonora.arnone@unito.it
[2] Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy
[3] MOX laboratory, Department of Mathematics, Politecnico di Milano, Milan, Italy

Abstract. Spatial regression models play a crucial role in analyzing environmental data and predicting spatially distributed phenomena. However, traditional approaches often struggle to capture the complex spatial dependencies and non-stationarities present in real-world datasets. In this paper, we propose a novel parameter cascading algorithm for spatial regression. The algorithm allows for the simultaneous estimation of the unknown spatial parameters describing the anisotropy and the spatial field itself, while incorporating physical and domain knowledge. We illustrate the proposed algorithm through an application to the analysis of rainfall data in Switzerland. The parameter cascading algorithm enables more accurate and localized predictions of spatially distributed variables.

Keywords: parameter cascading · partial differential equations · penalized regression · spatial data

1 Introduction

In many applications, especially in environmental sciences, capturing the spatial structure and variability of data is essential for making informed decisions and predictions. However, spatial data often exhibit complex patterns of anisotropy and non-stationarity, presenting challenges for traditional statistical modeling approaches.

Anisotropy, that is the directional dependence of spatial processes, is prevalent in many natural phenomena, such as those governed by flows, where the velocity field may exhibit different characteristics in different directions. In addition to anisotropy, many phenomena display non-stationarity. As an example, environmental variables like temperature or pollution levels may exhibit spatially varying trends or fluctuations [15].

Addressing strong forms of anisotropy and non-stationarity in spatial data requires to go beyond classical statistical methods to capture the complex spatial

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 77–81, 2025.
https://doi.org/10.1007/978-3-031-64431-3_13

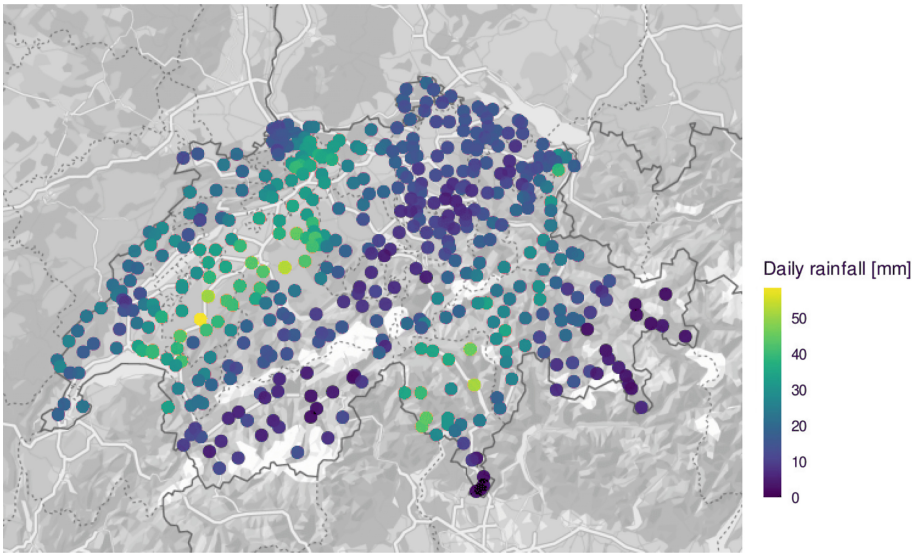


Fig. 1. Daily rainfall measurements recorded at 467 locations on Swiss territory on May 8, 1986. Bright colors correspond to higher values.

relationships and dynamics present in the data. A powerful mathematical tool to describe complex phenomena is given by differential equations. The integration of problem specific differential equations into spatial statistics methodologies has gained popularity due to the work of [4]. Other examples in this direction are found in [14,15], that include mechanistic information in dynamical spatio-temporal models. Another promising approach is physics-informed spatial regression, which integrates principles from physics, engineering, and related disciplines, into the statistical modeling framework. For example, [1] considers a parabolic, non-stationary Partial Differential Equation (PDE) in the estimation of a spatio-temporal blood velocity field. By incorporating physical laws, constraints, or governing equations, spatial regression methods can enhance the interpretability, accuracy, and robustness of the estimates.

This paper explores the application of the spatial regression with PDE regularization (SR-PDE) technique proposed in [2,3,11,12] for modeling anisotropy and non-stationarity in spatial data and propose a new method to estimate the anisotropy from the available data based on parameter cascading [10,13]. In particular, we apply the proposed method to rainfall data in Switzerland [7] where spatial variations are influenced by diverse geographical features and climatic factors. Figure 1 shows the considered rainfall data, highlighting the strong anisotropic pattern.

2 Parameter Cascading Algorithm

In the classical framework of SR-PDE (see, [3,11]) the anisotropy of a spatial field is modeled by the inclusion of a known PDE as a penalty in a regression functional. In this work, within this framework, instead of considering PDEs with known parameters, we deal with situations where the PDE in the regularization term contains unknown parameters. This allows us to fully leverage the modeling capabilities of PDEs, which are great for capturing complex phenomena behaviors.

To handle these unknown parameters in the PDE, we use the parameter cascading algorithm. This technique, originally devised for estimating parameters of differential equations [10], has been adapted to approximate parameters of PDEs. A first exploration of parameter cascading in SR-PDE has being done by [5], which has focused on the specific cases of stationary anisotropic diffusion. Our study broadens the scope, tackling more challenging cases involving general forms linear second-order elliptic PDEs with space-varying operators, allowing for possibly non-stationary PDE terms.

Essentially, parameter cascading entails the hierarchical arrangement of parameters in a model, wherein higher-level parameters influence lower-level ones. In our case, in addition to the classical parametric and non-parametric term in spatial regression model, called $\boldsymbol{\beta}$ and f respectively, we consider a second level of parameters, that are the one defining a second order linear elliptic PDE: the diffusion tensor K, the transport vector $\mathbf{b}$ and the reaction term c. Since the new objective functional is not convex in all these unknowns, a joint minimization is not feasible. Thus, an iterative algorithm based on parameter cascading is proposed.

Each iteration of the algorithm is composed by two steps. In the first step estimates of $\boldsymbol{\beta}$ and f are obtained by minimizing a regularized sum-of-squared error functional with respect to them, keeping fixed the estimates of K, $\mathbf{b}$, and c from the previous iteration. In the second step, using the newly obtained estimates of $\boldsymbol{\beta}$ and f, the parameters K, $\mathbf{b}$, and c are estimated iteratively by minimizing the classical sum-of-squared error with respect to each parameter separately. The sum-of-squared error depends on $\boldsymbol{\beta}$ and f computed in the previous step.

This iterative process allows for refining the estimates of the unknown PDE parameters, enhancing the accuracy of the model. Numerical optimization methods like Gradient Descent or Conjugate Gradient are employed to solve the minimization problems involved in estimating K, $\mathbf{b}$, and c. The initial values for these parameters are set to a default corresponding to isotropic smoothing unless specific prior knowledge is available.

3 Applications to Rainfall Data

We apply the proposed spatial regression method with the parameter cascading algorithm to analyze rainfall data in Switzerland. Switzerland's complex terrain and diverse climatic conditions present challenges for accurately modeling and predicting rainfall patterns across different regions.

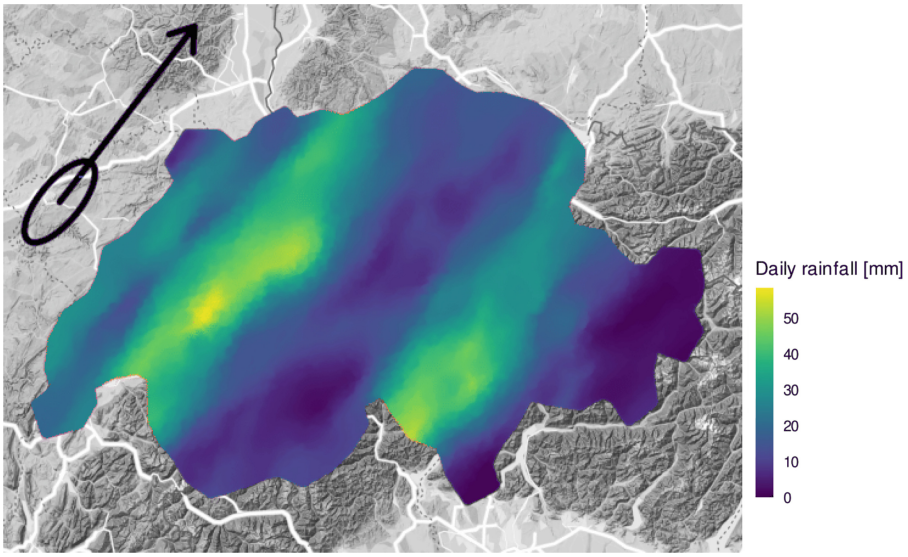


Fig. 2. Spatial field estimates of rainfall in Switzerland provided by the proposed method, with the corresponding estimated anisotropy ellipse and transport vector.

With the parameter cascading algorithm, researchers can capture the spatial heterogeneity of rainfall patterns across Switzerland, accounting for factors such as elevation, aspect, land cover, and local climate dynamics. This enables more accurate and localized predictions of rainfall intensity, duration, and frequency at unobserved locations, even in regions with sparse monitoring networks.

Figure 2 shows the obtained results: the diffusion ellipse and the transport vector represent quite well the underlying pattern in the data. This physical information estimated through parameter cascading is then added inside the SR-PDE model and it allows to achieve more precise results and predictions.

4 Conclusions

We presented a novel approach for including problem-specific differential equations into spatial statistics models, estimated through a parameter cascading algorithm. By leveraging the power of differential equations, we enhance the modeling flexibility and are able to capture complex spatial phenomena.

In the case of the fixed PDE parameters, uncertainty quantification within the considered models has been investigated in [6,8,9]. Developing appropriate inference procedure that also accounts for the uncertainty in the PDE parameters constitutes a fascinating direction for future research.

Acknowledgements. L.M. Sangalli acknowledges funding by the European Union - NextGenerationEU, research project PRIN2022 CoEnv - Complex Environmental Data and Modeling (CUP 2022E3RY23), and MUR research project Dipartimento di Eccellenza 2023–2027, Dipartimento di Matematica, Politenico di Milano.

References

1. Arnone, E., Azzimonti, L., Nobile, F., Sangalli, L.M.: Modeling spatially dependent functional data via regression with differential regularization. J. Multivar. Anal. **170**, 275–295 (2019)
2. Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P.: Mixed finite elements for spatial regression with PDE penalization. SIAM/ASA J. Uncertainty Quantification **2**(1), 305–335 (2014)
3. Azzimonti, L., Sangalli, L.M., Secchi, P., Domanin, M., Nobile, F.: Blood flow velocity field estimation via spatial regression with PDE penalization. J. Am. Stat. Assoc. **110**(511), 1057–1071 (2015)
4. Berliner, L.M.: Physical-statistical modeling in geophysics. J. Geophys. Res. Atmos. **108**(D24) (2003)
5. Bernardi, M.S., Carey, M., Ramsay, J.O., Sangalli, L.M.: Modeling spatial anisotropy via regression with partial differential regularization. J. Multivar. Anal. **167(C)**, 15–30 (2018)
6. Cavazzutti, M., Arnone, E., Ferraccioli, F., Finos, L., Galimberti, C., Sangalli, L.M.: A Novel Spatio-Sign-Flip tests for the nonparametric component in Spatial Regression with PDE regularization. In: Proceedings of the GRASPA 2023 Conference (2023)
7. Dubois, G., Malczewski, J., De Cort, M.J.M.R.: Mapping Radioactivity in the Environment: Spatial Interpolation Comparison 97. In: Office for Official Publications of the European Communities (2003) Switzerland rainfall data available at ftp.geog.uwo.ca/SIC97/intro/alldata.dat
8. Ferraccioli, F., Sangalli, L.M., Finos, L.: Some first inferential tools for spatial regression with differential regularization. J. Multivar. Anal. (2022). https://doi.org/10.1016/j.jmva.2021.104866
9. Ferraccioli, F., Sangalli, L.M., Finos, L.: Nonparametric tests for semiparametric regression models. TEST (2023). https://doi.org/10.1007/s11749-023-00868-9
10. Ramsay, J.O., Hooker, G., Campbell, D., Cao, J.: Parameter estimation for differential equations: a generalized smoothing approach. J. Roy. Stat. Soc. Series B (Stat. Methodol.) **69**(5), 741–796 (2007)
11. Sangalli, L.M.: Spatial regression with partial differential equation regularisation. Int. Stat. Rev. **89**, 505–531 (2021)
12. Sangalli, L.M., Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. J. Roy. Stat. Soc. B **75**, 681–703 (2013)
13. Xun, X., Cao, J., Mallick, B., Maity, A., Carroll, R.J.: Parameter estimation of partial differential equation models. J. Am. Stat. Assoc. **108**(503), 1009–1020 (2013)
14. Wikle, C.K.: Hierarchical bayesian models for predicting the spread of ecological processes. Ecology **84**(6), 1382–1394 (2003)
15. Wikle, C.K., Hooten, M.: A general science-based framework for dynamical spatio-temporal models. TEST **19**, 417–451 (2010)



A New Distribution for Overdispersed Count Data

Roberto Ascari[1(✉)], Agnese M. Di Brisco[2], and Sonia Migliorati[1]

[1] Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Milan, Italy
{roberto.ascari,sonia.migliorati}@unimib.it
[2] Department of Studies for Economics and Business (DISEI), University of Piemonte Orientale, Novara, Italy
agnese.dibrisco@uniupo.it

Abstract. This work aims to introduce a novel distribution for count data which is capable of accounting for issues such as overdispersion and an excess of zeros. Currently, the reference distributions for handling such data are the negative binomial (NegBin) and the beta-negative binomial (BNB). The proposed distribution, named the flexible beta-negative binomial (FBNB), exhibits a compound structure that resembles that of the BNB while increasing flexibility in capturing heavy tails, excess of zeros, and latent structures in the data. The paper illustrates the characterization of the FBNB as a special finite mixture with two BNB components, shows the moments of the FBNB, and provides insights into the interpretation of its parameters. In addition, a brief introduction of a regression model aimed at evaluating the impact of covariates on the component-specific means is provided. It is worth noting that the results concerning the estimation of unknown parameters for both the FBNB and the associated regression model are still in the preliminary stages and, as such, are not presented in this paper.

Keywords: negative binomial · compound approach · mixture model · overdispersion

1 Introduction

When dealing with count data, a common choice is to use some statistical method resorting to the Poisson distribution [1]. Despite its simplicity, it is well-known that the Poisson distribution imposes the equality of the mean and variance, a condition referred to as equidispersion. However, it is common for real datasets to be overdispersed, namely having a variance that exceeds the mean. The overdispersion issue may arise due to several reasons, including the presence of latent clusters or an excess of zeros (i.e., a higher proportion of zero counts than the one expected by the Poisson distribution). Proper distributions to account for overdispersion are the NegBin and the BNB [5]. This work aims to propose a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 82–87, 2025.
https://doi.org/10.1007/978-3-031-64431-3_14

novel distribution for overdispersed count data characterized by increased flexibility. The proposed model, named FBNB, is obtained by a compound approach as the BNB.

Indeed, the BNB is obtained by compounding the NegBin distribution, assuming that its "success probability" parameter Π follows a beta distribution. Similarly, the FBNB is obtained by assuming that Π follows a FB distribution [4]. The FB is defined as a special identifiable mixture with two betas and greatly extends the shapes of the beta probability density function p.d.f. mainly in terms of asymmetry, bimodality, and heavy tail behavior. As a result, these characteristics are transferred to the FBNB, thus enhancing its flexibility with respect to the NegBin and the BNB distributions.

This paper aims to provide a preliminary study on the proposed distribution, exploiting its special mixture structure with BNB-distributed components, its first two moments, and the interpretation of the parameters in terms of their effect on the probability mass function p.m.f. Finally, the paper sketches a regression model taking advantage of the FBNB's mixture structure by regressing onto covariates its component-specific means. Simulation studies concerning the estimation of unknown parameters for both the FBNB and the associated regression model are in progress and, as such, are deferred to a further paper.

The outline of this work is as follows. Section 2 provides a review of the NegBin and BNB distributions, while Sect. 3 introduces the novel FBNB distribution. Section 4 briefly discusses the regression model built upon the FBNB and touches upon the remaining open issues.

2 The Negative Binomial and the Beta-Negative Binomial

The NegBin is a widespread alternative to the Poisson distribution for count data exhibiting overdispersion. Let Y be a discrete random variable (r.v.) following a NegBin(k, π) distribution. Then, its p.m.f is defined as:

$$f_{\text{NB}}(y|k,\pi) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}(1-\pi)^y \pi^k,$$

where $y \in \mathbb{N}$, $k > 0$, $\pi \in (0,1)$, and $\Gamma(\cdot)$ indicates the Gamma function. The parameter k is, in general, a real positive value, but when $k \in \mathbb{N}$ (i.e., it is an integer), it allows for an intuitive interpretation. Indeed, when $k \in \mathbb{N}$, a NegBin-distributed r.v. models the number of failures before k successes, each having success probability π.

The NegBin does not impose the equidispersion condition, therefore being more flexible than the Poisson, but not flexible enough to account for a heavy right-hand tail. On the contrary, the BNB further extends the flexibility of the NegBin by accounting for both overdispersed data and a heavy right-hand tail. The BNB can be introduced by compounding the NegBin distribution as follows. Let $Y|\Pi = \pi$ be a discrete r.v. following a NegBin (k, π) distribution and $\Pi \sim$

Beta(α_1, α_2), where $k > 0$, $\alpha_1 > 0$, and $\alpha_2 > 0$. Then, Y is marginally distributed as a BNB(k, α_1, α_2) random variable with p.m.f.

$$f_{\text{BNB}}(y|k, \alpha_1, \alpha_2) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \frac{B(\alpha_1 + k, \alpha_2 + y)}{B(\alpha_1, \alpha_2)}, \tag{1}$$

where $y \in \mathbb{N}$, $k > 0$, $\alpha_1 > 0$, $\alpha_2 > 0$, and $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ represents the Beta function.

The parameter k can be thought of as a dispersion parameter, whereas the parameters α_1 and α_2 control the heaviness of the right-hand tail and the location, respectively [2]. In particular, the first two order moments of a BNB distribution can be expressed as:

$$\mathbb{E}[Y] = k\frac{\alpha_2}{\alpha_1 - 1},$$

$$\text{Var}(Y) = k\frac{\alpha_2(k + \alpha_1 - 1)(\alpha_1 + \alpha_2 - 1)}{(\alpha_1 - 2)(\alpha_1 - 1)^2}.$$

It is worth noting that the mean and variance exist finite only for $\alpha_1 > 1$ and $\alpha_1 > 2$, respectively. Finally, an issue concerning the BNB is that it is not identifiable, namely it is possible to swap places of parameters k and α_2 while still obtaining the same p.m.f..

3 The Flexible Beta-Negative Binomial Distribution

The FBNB arises as a compound distribution where $Y|\Pi = \pi \sim \text{NegBin}(k, \pi)$, and Π follows an FB distribution, $\Pi \sim \text{FB}(\alpha_1, \alpha_2, \tau, p)$ [4]. It is useful to recall that the p.d.f. of the FB is as follows:

$$f_{\text{FB}}(\pi | \alpha_1, \alpha_2, \tau, p) = \\ = \frac{\Gamma(\alpha_1 + \alpha_2 + \tau)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1} \left[p\frac{\Gamma(\alpha_1)\pi^{\tau}}{\Gamma(\alpha_1 + \tau)} + (1-p)\frac{\Gamma(\alpha_2)(1-\pi)^{\tau}}{\Gamma(\alpha_2 + \tau)} \right],$$

where $\pi \in (0, 1)$, $\alpha_1 > 0$, $\alpha_2 > 0$, $\tau > 0$, and $p \in (0, 1)$. The FB can be written as a mixture with two beta components, where the parameter τ controls the distance between the expectations of the two components (i.e., the higher τ, the higher the distance). For further details on the FB please refer to [4].

The resulting marginal distribution of Y is thus FBNB with the following p.m.f.:

$$f_{\text{FBNB}}(y | k, \alpha_1, \alpha_2, \tau, p) = \\ = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left[p\frac{B(\alpha_1 + \tau + k, \alpha_2 + y)}{B(\alpha_1 + \tau, \alpha_2)} + (1-p)\frac{B(\alpha_1 + k, \alpha_2 + \tau + y)}{B(\alpha_1, \alpha_2 + \tau)} \right],$$

where $y \in \mathbb{N}$, $k > 0$, $\alpha_1 > 0$, $\alpha_2 > 0$, $\tau > 0$, and $p \in (0, 1)$. Interestingly, the FBNB can be interpreted as a mixture with two BNB distributed components, namely:

$$f_{\text{FBNB}}(y|k, \alpha_1, \alpha_2, \tau, p) = p f_{\text{BNB}}(y|k, \alpha_1 + \tau, \alpha_2) + (1-p) f_{\text{BNB}}(y|k, \alpha_1, \alpha_2 + \tau).$$

The first and second BNB-distributed components have expected values equal to $\lambda_1 = k\dfrac{\alpha_2}{\alpha_1+\tau-1}$ and $\lambda_2 = k\dfrac{(\alpha_2+\tau)}{\alpha_1-1}$, respectively. As a result, the marginal expectation of an FBNB-distributed r.v. can be obtained in light of its mixture structure as $\mathbb{E}[Y] = p\lambda_1 + (1-p)\lambda_2$, which exists finite only for $\alpha_1 > 1$.

Given $\alpha_1 > 1$, it is also possible to show that

$$\lambda_2 - \lambda_1 = k\tau(\alpha^+ + \tau - 1)/((\alpha_1 - 1)(\alpha_1 + \tau - 1)) > 0,$$

where $\alpha^+ = \alpha_1 + \alpha_2$. This result guarantees that a constraint exists between the component means, being $\lambda_2 > \lambda_1$, thus excluding any label switching issue due to the invariance under relabeling of the mixture components.

By taking advantage of the above-highlighted mixture structure, it is possible also to derive the expression for the variance, namely

$$\begin{aligned}\mathrm{Var}(Y) = p\lambda_1 &\left[\frac{(\alpha^+ + \tau - 1)(\alpha_1 + \tau + k - 1)}{(\alpha_1 + \tau - 2)(\alpha_1 + \tau - 1)} + \lambda_1\right] + \\ &+ (1-p)\lambda_2\left[\frac{(\alpha^+ + \tau - 1)(\alpha_1 + k - 1)}{(\alpha_1 - 2)(\alpha_1 - 1)} + \lambda_2\right] - [p\lambda_1 + (1-p)\lambda_2]^2.\end{aligned}$$

Please note that the variance of the FBNB exists and is finite only for $\alpha_1 > 2$.

Inspecting Figs. 1, 2, and 3, it is possible to study how the parameters p, τ, and k affect the FBNB pmf and thus how to interpret them. As p increases (Fig. 1), it is worth noting that most of the mass is shifted towards the smaller values of y. This result is coherent with the observation that parameter p is the mixing weight of the first mixture component and that the first component has a smaller expected value than the second one since $\lambda_1 < \lambda_2$.

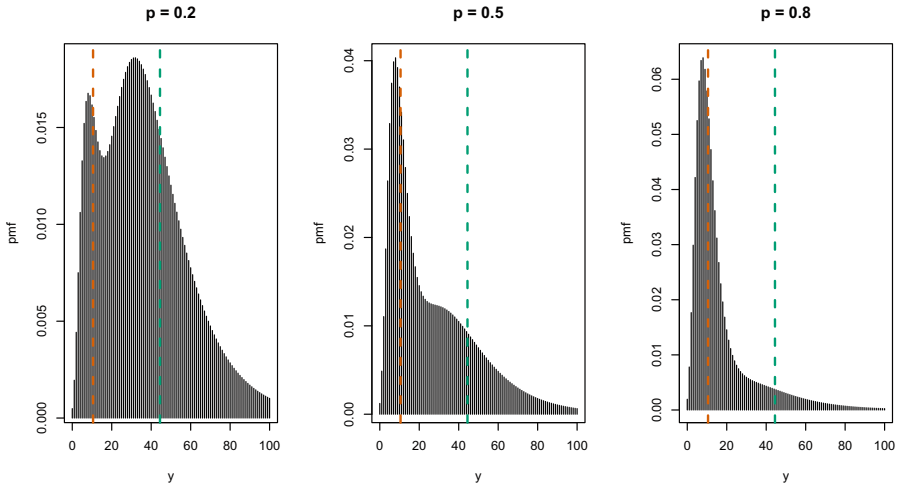


Fig. 1. p.m.f. of FBNB on a subset of its support for $\alpha_1 = \alpha_2 = 10$, $k = 20$, $\tau = 10$, and $p \in \{0.2, 0.5, 0.8\}$. Dashed lines intersect the points λ_1 (orange line) and λ_2 (green line). (Color figure online)

The parameter τ has the same impact on the FBNB distribution as it has on the FB distribution since it controls the distance between the mixture components. Indeed, as τ increases (Fig. 2), the two component-specific means become more distant, enhancing the bimodality and heaviness of the right-hand tail.

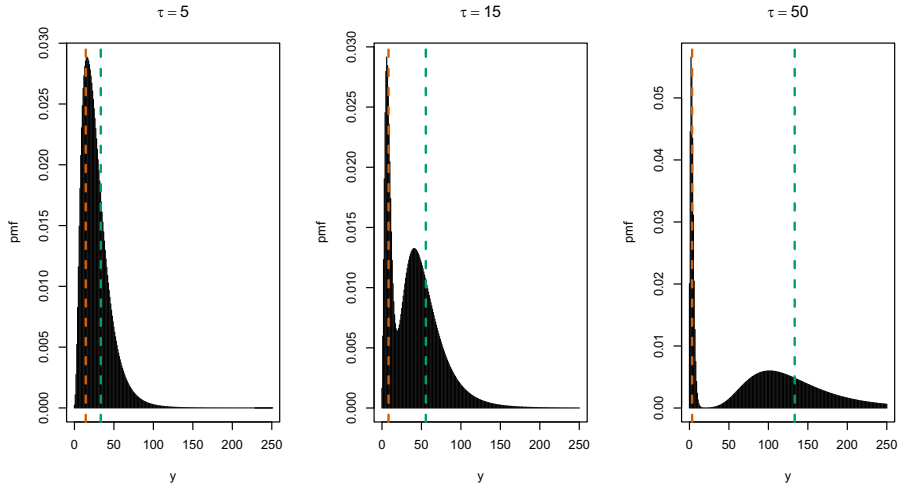


Fig. 2. p.m.f. of FBNB on a subset of its support for $\alpha_1 = \alpha_2 = 10$, $k = 20$, $p = 0.3$, and $\tau \in \{5, 15, 50\}$. Dashed lines intersect the points λ_1 (orange line) and λ_2 (green line). (Color figure online)

The parameter k can be thought of as a dispersion parameter, as clear from the third panel of Fig. 3. Indeed, smaller values of k are associated with a smaller variance and a mass that is shifted towards small values of y.

Finally, due to space constraints, there is no room for a formal proof, but intuitively the FBNB, being a mixture with non-identifiable components, is non-identifiable itself.

4 The Flexible Beta-Negative Binomial Regression Model

Let us consider a vector of independent responses $Y_1, \ldots, Y_n$, where $Y_i \in \mathbb{N}, i = 1, \ldots, n$, where each $Y_i \sim \text{FBNB}(k, \alpha_1, \alpha_2, \tau, p)$. By adopting a GLM-type methodology [3], we propose a regression model on the two component-specific means that can be defined as

$$g(\lambda_{1i}) = \mathbf{x}_i^\intercal \boldsymbol{\beta}_1,$$
$$g(\lambda_{2i}) = \mathbf{z}_i^\intercal \boldsymbol{\beta}_2.$$

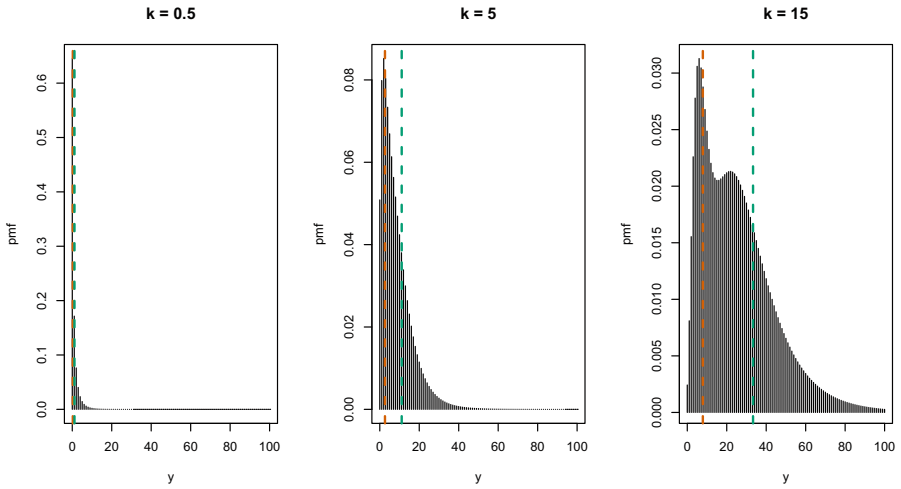


Fig. 3. p.m.f. of FBNB on a subset of its support for $\alpha_1 = \alpha_2 = 10$, $\tau = 10$, $p = 0.3$, and $k \in \{0.5, 5, 15\}$. Dashed lines intersect the points λ_1 (orange line) and λ_2 (green line). (Color figure online)

Please note that $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of possibly overlapping covariates, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of regression coefficients, and $g(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is an adequate link function, such as the logarithm.

It is worth noting that the regression model as it is does not take into account the constraint existing among the parameters, being $\lambda_2 > \lambda_1$. A further specification is thus needed to properly handle this aspect. Some proposals to cope with this and the previously stated open issues are currently under study and will be illustrated in future works.

References

1. Davis, R.A., Dunsmuir, W.T., Streett, S.B.: Observation-driven models for Poisson counts. Biometrika **90**(4), 777–790 (2003)
2. Gorgi, P.: Beta-negative binomial auto-regressions for modelling integer-valued time series with extreme observations. J. R. Stat. Soc. Ser. B Stat Methodol. **82**(5), 1325–1347 (2020)
3. McCullagh, P., Nelder, J.A.: Generalized Linear Models, Chapman & Hall (1989)
4. Migliorati, S., Di Brisco, A.M., Ongaro, A: A new regression model for bounded responses. Bayesian Anal. **13**(3), 845-872 (2018)
5. Wang, Z.: One mixed negative binomial distribution with application. J. Stat. Plann. Infer. **141**(3), 1153–1160 (2011)



Uncertainty Reduction in a Class of Dependent Dirichlet Processes

Filippo Ascolani[1(✉)], Stefano Damato[2], and Matteo Ruggiero[3]

[1] Duke University, 214 Old Chemistry 27708, NC, USA
filippo.ascolani@duke.edu
[2] IDSIA USI-SUPSI, Via La Santa 1, Lugano 6962, Switzerland
stefano.damato@idsia.ch
[3] University of Torino and Collegio Carlo Alberto, C.so Unione Sovietica 218/bis, 10134 Turin, Italy
matteo.ruggiero@unito.it

Abstract. We consider a dependent Dirichlet process model driven by a Fleming–Viot diffusion, with observations collected at discrete time points in a Hidden Markov model fashion. We investigate empirically a phenomenon described in [2], whereby upon conditioning the underlying random probability measure on data collected at past and future times, the mixture distribution which describes the posterior distribution of the random measure benefits from an automatic reduction of the number of components. This reduction depends on the observed multiplicities and upweights components which feature observations that are shared across multiple times, and is bound to have positive implications for inference in terms of reduction of the estimation uncertainty and computational cost.

Keywords: Hidden Markov model · smoothing · mixture · Bayesian nonparametrics

1 Introduction

Let $X := \{X_t, t \geq 0\}$ be a Fleming–Viot driven Dependent Dirichlet process (FV-DDP) (cf. [1,2,9]) with baseline measure $\alpha = \theta P_0$, where $\theta > 0$ and P_0 a probability measure on an arbitrary Polish space $\mathbb{X}$, and reversible measure given by the law Π_α of a Dirichlet process [5]. Formally, the FV-DDP is a diffusion process with values on the space of purely atomic probability measures on $\mathbb{X}$, whose time-homogeneous dynamics can be described as

$$X_t | m, Y_1, \ldots, Y_m \sim \Pi_{\alpha + \sum_{i=1}^m \delta_{Y_i}}, \qquad Y_i \overset{\text{iid}}{\sim} X_0, \tag{1}$$

where $m \in \mathbb{Z}_+$ is the value at t of a suitable death process which start from ∞ at time 0. Cf. [1] for this formulation. See also [3,4] for generalities on Fleming–Viot processes.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 88–92, 2025.
https://doi.org/10.1007/978-3-031-64431-3_15

We assume observations are collected at discrete times $0 = t_0 < t_1 < \ldots, t_{p-1} = T$, not necessarily equally spaced, and denote with $Y_{0:T} = (Y_0, \ldots, Y_T)$ the overall dataset, where each Y_t denotes a vector of observations. The full statistical model is described by the following hidden Markov model

$$Y_{t,i} \mid X_t = x \overset{\text{iid}}{\sim} x, \quad X \sim \text{FV-DDP}(\alpha), \quad X_0 \sim \Pi_\alpha, \tag{2}$$

where $X \sim \text{FV-DDP}(\alpha)$ denotes that X is as in (1), and the initial value X_0 is a Dirichlet process with parameter α. Thus, at each time t observations are assumed to be conditionally independent and sampled from the realization of X_t. This implies that if $X_t \overset{d}{=} \sum_{j\geq 1} W_j \delta_{Z_j}$, then $Y_{t,i} = Z_j$ with probability W_j. Therefore (2) extends to a continuous-time framework the classical Dirichlet process model for exchangeable observations.

Under such model specification, [2] provided a full description of the conditional distribution of X_t, given $Y_{0:T}$, which we recall here in the simplified scenario where $Y_{0:T} = (Y_0, Y_t, Y_T)$, namely only a single past and future collection time (with respect to time t) are available. In the hidden Markov models literature this corresponds to the so-called *smoothing distribution*, which combines past and future information to improve a previously obtained estimate of X_t given only past data.

Let $\mathbf{y}^* = (y_1^*, \ldots, y_K^*)$, with $K \in \mathbb{N}$, denote the vector of distinct values observed in $Y_{0:T}$. We let $\mathbf{n}_t = (n_{t,1}, \ldots, n_{t,K}) \in \mathbb{Z}_+^K$ denote the multiplicities of the distinct types in $\mathbf{y}^*$ observed in Y_t. We also set $|\mathbf{n}| := \sum_{k=1}^K n_k$ and we say that $\mathbf{m} \leq \mathbf{n}$ if $m_k \leq n_k$ for every k. Then [2] showed that such conditional distribution is

$$X_t | \, Y_{0:T}, \mathbf{m}_0, \mathbf{m}_T \sim \Pi_{\alpha + \sum_{k=1}^K (m_{0,k} + n_{t,k} + m_{T,k})\delta_{y_k^*}}, \tag{3}$$

where $\mathbf{n}_t$ is given, while $\mathbf{m}_0, \mathbf{m}_T$ are latent variables on

$$\{(\mathbf{m}_0, \mathbf{m}_T) \in \mathbb{Z}_+^2 : \mathbf{0} \leq \mathbf{m}_0 \leq \mathbf{n}_0, \mathbf{0} \leq \mathbf{m}_T \leq \mathbf{n}_T\}$$

with distribution

$$p(\mathbf{m}_0, \mathbf{m}_T | Y_{0:T}) \propto p_{\mathbf{n}_0, \mathbf{m}_0}(t) p_{\mathbf{n}_T, \mathbf{m}_T}(T - t) K(\mathbf{m}_0, \mathbf{n}_t, \mathbf{m}_T). \tag{4}$$

In the latter, $p_{\mathbf{n},\mathbf{m}}(s)$ is the transition probability of a death process on $\mathbb{Z}_+^K$ which starts from $\mathbf{n}$ and jumps from $\mathbf{k}$ to $\mathbf{k} - \mathbf{e}_i$ at rate $k_i(\theta + |\mathbf{k}| - 1)/2$, where we have denoted by $\mathbf{e}_i$ the canonical vector in the i-th direction. See [1,6–9] for special cases and variants of this problem. The specific formula for $K(\mathbf{m}_0, \mathbf{n}_t, \mathbf{m}_T)$, which is an aggregation term combining past, present and future observations, depends crucially on the baseline distribution (see Proposition 3.6 in [2]).

The above expressions imply that $X_t | Y_{0:T}$ has law given by a finite mixture of laws of Dirichlet processes. However, it can be shown that the cardinality of the mixture grows polynomially with the number of collected observations (see also Section 3 in [2]). Thus the computational cost of computing the smoothing

distribution may become unfeasible even with datasets of moderate sizes, which requires to devise careful numerical approximations (as done in [6] for the finite-dimensional counterpart of model (2)).

In this paper we explore the impact of choosing P_0 to be a nonatomic probability measure from a computational viewpoint. We will show that this specification implies an automatic pruning of the finite mixture in (3), when some observations are shared between past and future collection times.

2 Uncertainty Reduction in Fleming–Viot Smoothing

Under the above model specification, assume that some unique values in $\mathbf{y}^*$ are observed multiple times in 0 and T, i.e. defining

$$D_0 = \{j \in \{1, \dots, K\} \mid n_{0,j} > 0 \text{ and either } n_{t,j} > 0 \text{ or } n_{T,j} > 0\}$$

and

$$D_T = \{j \in \{1, \dots, K\} \mid n_{T,j} > 0 \text{ and either } n_{t,j} > 0 \text{ or } n_{0,j} > 0\},$$

then the set

$$D = \left\{(\mathbf{k}, \mathbf{k}') \leq (\mathbf{n}_0, \mathbf{n}_T) \mid \mathbf{k}_j > 0, \mathbf{k}'_{j'} > 0 \text{ for every } j \in D_0 \text{ and } j' \in D_T\right\} \quad (5)$$

is not empty. In words, D contains the nodes (i.e. combinations of past and future observations) whose frequency of unique values observed in distinct collection times is non null.

In this setting [2] showed that, if a non-atomic probability measure P_0 is chosen as the baseline distribution in (2), the aggregation term (4) becomes

$$K(\mathbf{m}_0, \mathbf{n}_t, \mathbf{m}_T) = \frac{\theta_{(|\mathbf{m}_{0,k}|)}\theta_{(|\mathbf{m}_{T,k}|)}}{(\theta + |\mathbf{n}_t|)_{(|\mathbf{m}_{0,k}|+|\mathbf{m}_{T,k}|)}} \prod_{j=1}^{K} \frac{(m_{0,j} + n_{t,j} + m_{T,j} - 1)!}{(m_{0,j} - 1)!(n_{t,j} - 1)!(m_{T,j} - 1)!} \quad (6)$$

if $(\mathbf{m}_0, \mathbf{m}_T) \in D$ and zero elsewhere, where we denote by $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$ the ascending factorial or Pochhammer symbol and set $(-1)! = 1$ by convention. Thus, the smoothing distribution reduces nodes with positive probability from those in the set $\{(\mathbf{k}, \mathbf{k}') \leq (\mathbf{n}_0, \mathbf{n}_T)\}$ to those in D.

We illustrate this effect in Fig. 1 when $K = 4$ and

$$\mathbf{n}_0 = (2, 1, 1, 0), \quad \mathbf{n}_t = (1, 0, 1, 0), \quad \mathbf{n}_T = (3, 0, 0, 1). \quad (7)$$

Here the first type is shared across all three observation times, while the third one only between the past and present times. In the figure, the points along the x-axis are values of $\mathbf{m}_0$, and those along the y-axis of $\mathbf{m}_T$. The graphics shows that only nodes $(\mathbf{m}_0, \mathbf{m}_T) \leq (\mathbf{n}_0, \mathbf{n}_T)$ such that $m_{0,1} > 0$, $m_{0,3} > 0$ and $m_{T,1} > 0$ are assigned positive probability (blue balls), while all others have null mass (red crosses). Hence, a nonatomic P_0 implies a massive reduction in

the computational burden, since the majority of nodes (hence of mixture components) that in principle must be considered for drawing inferences, can in practice be discarded. Moreover, by reducing the number of elements in the mixture, the uncertainty quantified by the smoothing distribution is significantly reduced. Finally, Fig. 1 also highlight empirically that the aggregation term (6) favours nodes with shared information: for example, the combination with the higher weight corresponds to $\mathbf{m}_0 = (3, 0, 0, 0)$ and $\mathbf{m}_T = (2, 0, 1, 0)$, which maximizes the frequency of the shared values and minimizes the ones of types observed only at a single collection time.

If instead we opt for a different specification of the baseline distribution P_0, e.g., chosen to be discrete with a finite support, the above described model falls under the special case of a Wright–Fisher hidden Markov model. In such case, the observation of multiplicities as in (7) would instead imply that all the nodes in the picture would have positive probability.

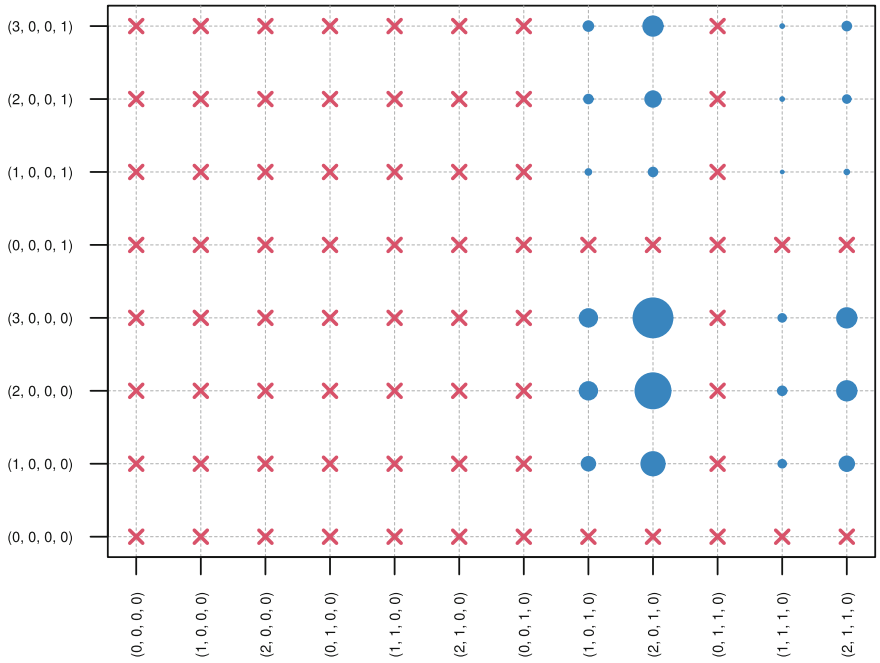


Fig. 1. Depiction of weights (6) with $\mathbf{m}_0$ on the x-axis and $\mathbf{m}_T$ on the y-axis, where $(\mathbf{n}_0, \mathbf{n}_t, \mathbf{n}_T)$ is as in (7). The red crosses represent nodes with zero probability, while the size of the blue balls is proportional to the corresponding weight. (Color figure online)

3 Conclusion

We show empirically that choosing a non-atomic baseline distribution in model (2) leads to some desirable properties, as it reduces the cardinality of the mixture which represents the smoothing distribution (or marginal posterior) of the unobserved process state, in turn drastically reducing the implied computational cost for computing it and the uncertainty induced in terms of inference. It is interesting to notice that this phenomenon vanishes by choosing an atomic P_0, see Proposition 3.6 in [2].

The set D in (5) corresponds to the possible combinations which entail positive probability. Even if its cardinality is much smaller than the set of nodes which in principle are assigned positive probability, by virtue of the aforementioned pruning effect, it may not be a trivial task to enumerate its elements, especially when K is high. This issue will be investigated in future works.

References

1. Ascolani, F., Lijoi, A., Ruggiero, M.: Predictive inference with Fleming-Viot-driven dependent Dirichlet processes. Bayesian Anal. **16**, 371–395 (2021)
2. Ascolani, F., Lijoi, A., Ruggiero, M.: Smoothing distributions for conditional Fleming-Viot and Dawson-Watanabe diffusions. Bernoulli **29**, 1410–1434 (2023)
3. Ethier, S.N., Griffiths, R.C.: The transition function of a Fleming-Viot process. Ann. Probab. **21**, 1571–1590 (1993)
4. Ethier, S.N., Kurtz, T.G.: Fleming-Viot processes in population genetics. SIAM J. Control. Optim. **31**, 345–386 (1993)
5. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)
6. Kon Kam King, G., Papaspiliopoulos, P., Ruggiero, M.: Exact inference for a class of hidden Markov models on general state spaces. Electr. J. Stat. **15**, 2832–2875 (2021)
7. Kon Kam King, G., Pandolfi, A., Piretto, M., Ruggiero, M.: Approximate filtering via discrete dual processes. Stoch. Proc. Appl. **168**, 104268 (2024)
8. Papaspiliopoulos, O., Ruggiero, M.: Optimal filtering and the dual process. Bernoulli **20**, 1999–2019 (2014)
9. Papaspiliopoulos, O., Ruggiero, M., Spanò, D.: Conjugacy properties of time-evolving Dirichlet and gamma random measures. Electron. J. Stat. **10**, 3452–3489 (2016)

Non-Informative Priors in Wright–Fisher Smoothing

Filippo Ascolani[1], Ylenia F. Buttigliero[2(✉)], and Matteo Ruggiero[2]

[1] Duke University, 214 Old Chemistry, Durham 27708, NC, USA
filippo.ascolani@duke.edu
[2] University of Torino and Collegio Carlo Alberto, C.so Unione Sovietica 218/bis, 10134 Turin, Italy
{yleniafrancesca.buttigliero,matteo.ruggiero}@unito.it

Abstract. We consider a hidden Markov model for discretely observed binary data, with underlying unobserved dynamic probabilities driven by a one-dimensional Wright–Fisher diffusion. We leverage on recent results on the posterior distribution of the diffusion state given data collected at two time points, to investigate a non-informative prior specification for inference at an intermediate time. Our findings describe explicitly the probability distribution of the data points retention for inference at this intermediate time .

Keywords: Hidden Markov model · smoothing density · non-informative prior

1 Introduction

Assume we collect binary observations $\boldsymbol{Y}_t = (Y_{t,1}, \dots, Y_{t,n_t})$ at different times, not necessarily regularly spaced, according to the following hidden Markov model:

$$Y_{t,i} | X_t = x \overset{\text{iid}}{\sim} \text{Bern}(x), \quad X \sim \text{WF}(\alpha_1, \alpha_2), \quad \alpha_i > 0, \tag{1}$$

Here $\text{Bern}(x)$ denotes a Bernoulli distribution with success probability x. The temporal dependence between observations, which are independent only given the current X_t, is modelled by assuming $X = \{X_t, t \geq 0\}$ is an unobserved diffusion process (*signal*) which evolves like a Wright–Fisher diffusion on $[0, 1]$ with infinitesimal generator

$$Lf(x) = \frac{1}{2}\left[\alpha_1(1-x) - \alpha_2 x\right] f'(x) + \frac{1}{2} x(1-x) f''(x), \tag{2}$$

acting on functions $C^2([0, 1])$. The Wright–Fisher was introduced in population genetics to model the evolution of the frequency of a particular allele at a single locus in a large panmictic population, subjected to random mating and parent-independent mutation. In this framework it can however be simply interpreted as the time-varying probability of success in dynamic coin toss problem. These

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 93–97, 2025.
https://doi.org/10.1007/978-3-031-64431-3_16

have been recently considered in a statistical framework such as (1) in [4,5,7]. See also [1,2,8] for nonparametric extensions.

The Wright–Fisher diffusion is reversible and stationary with respect to the Beta distribution $\text{Beta}(\alpha_1, \alpha_2)$ with density on $[0,1]$ proportional to $x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}$. Thus, letting $X_0 \sim \text{Beta}(\alpha_1, \alpha_2)$ at each time t the marginal model induced by (1) is given by

$$Y_{t,i} | X_t = x \overset{\text{iid}}{\sim} \text{Bernoulli}(x), \quad X_t \sim \text{Beta}(\alpha_1, \alpha_2). \tag{3}$$

Therefore, (1) can be seen as a dynamic counterpart of the classical Beta-Bernoulli model for binary data. Within this perspective, the mutation rates α_i play the role of the prior parameters which specify the strength of our prior information. Indeed, the classical conjugacy for Beta-Bernoulli models (in our present notation) reads

$$X_t \mid \boldsymbol{Y}_t \sim \text{Beta}\left(\alpha_1 + \sum_{i=1}^{n_t} Y_{t,i}, \alpha_2 + n_t - \sum_{i=1}^{n_t} Y_{t,i}\right), \tag{4}$$

so that α_1 (resp. α_2) is the amount of fictitious observations equal to 1 (resp. 0). It is clear from (4) that the higher α_1 and α_2 the higher the impact of the prior, relative to the data, on the posterior distribution. It is thus of interest to understand what happens when no prior information is available, i.e. $\alpha_1 \to 0$ and/or $\alpha_2 \to 0$. This is well known in the static setting [9], while here we extend this analysis to the corresponding temporal specification as in (1).

We assume to observe two samples $\boldsymbol{Y}_0$ and $\boldsymbol{Y}_T$ at times 0 and $T > 0$ respectively, whose multiplicities of observed 0's and 1's are $\mathbf{n}_0 = (n_{0,0}, n_{0,1})$ and $\mathbf{n}_T = (n_{T,0}, n_{T,1})$ respectively. Moreover, for $\mathbf{m}, \mathbf{n} \in \mathbb{Z}_+^2$ with $\mathbf{m} = (m_0, m_1)$ and similarly for $\mathbf{n}$, we say that $\mathbf{m} \leq \mathbf{n}$ if $m_0 \leq n_0$ and $m_1 \leq n_1$ and $\mathbf{m} < \mathbf{n}$ if $\mathbf{m} \leq \mathbf{n}$ and at least one component of $\mathbf{m}$ is strictly lower than the corresponding component of $\mathbf{n}$, i.e. $m_0 < n_0$ and/or $m_1 < n_1$.

We are interested in the posterior distribution of the signal at time t, with $0 < t < T$, i.e. at an intermediate time between the two collected samples. In the Hidden Markov model literature such objects, which combine past and future observations, are called *smoothing* distributions. Recent works [4] showed that the law under $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ for X_t given observations as described above is

$$\mathcal{L}_{\boldsymbol{\alpha}}(X_t | \boldsymbol{Y}_0, \boldsymbol{Y}_T) = \sum_{\mathbf{n} \leq \mathbf{n}_0} \sum_{\mathbf{m} \leq \mathbf{n}_T} W^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}(t;T) \text{Beta}(\alpha_1 + m_1 + n_1, \alpha_2 + m_0 + n_0), \tag{5}$$

where the mixing weights are given by

$$W^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}(t;T) \propto p^{\boldsymbol{\alpha}}_{\mathbf{n}_0,\mathbf{n}}(t) p^{\boldsymbol{\alpha}}_{\mathbf{n}_T,\mathbf{m}}(T-t) K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}, \tag{6}$$

with $p^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$ the transition probability of a death process (see (7) below) and $K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$ as in (10). In words, the smoothing distribution places positive mass on all the combinations $(\mathbf{n}, \mathbf{m})$ which are the possible arrival points of two independent death processes started from $\mathbf{n}_0$ and $\mathbf{n}_T$ respectively. The weights assigned to each combination are given by the product of the transition probabilities of such death processes and an aggregation term $K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$. See [4] for more details.

2 Smoothing Under Non-Informative Priors

From expression (5) we appreciate that the dependence on $\boldsymbol{\alpha}$ is both through the mixing weights and the Beta components. In the following paragraphs we study separately each term when $\alpha_1 \to 0$ and $\alpha_2 \to 0$, which corresponds to a non-informative setting.

2.1 The Transition Probabilities

It is well known that the Wright–Fisher diffusion in (2) is dual to Kingman's coalescent (cf., e.g., [7]), which is a pure death process on $\mathbb{Z}_+^2$. Its transition probabilities can be written

$$p^{\boldsymbol{\alpha}}_{\mathbf{m},\mathbf{m}'}(t) = \begin{cases} e^{-\lambda^{\boldsymbol{\alpha}}_{|\mathbf{m}|}t} & \text{if } \mathbf{m}' = \mathbf{m} \\ C^{\boldsymbol{\alpha}}_{|\mathbf{m}|,|\mathbf{m}'|}(t)p(\mathbf{m},\mathbf{m}') & \text{if } \mathbf{m}' < \mathbf{m} \end{cases}, \tag{7}$$

where $|\mathbf{m}| = m_0 + m_1$, $C^{\boldsymbol{\alpha}}_{|\mathbf{m}|,|\mathbf{m}'|}(t)$ is the transition probability of a inhomogeneous death process on $\mathbb{N}$ with rates $\lambda^{\boldsymbol{\alpha}}_{|\mathbf{m}|} = \frac{|\mathbf{m}|(|\boldsymbol{\alpha}|+|\mathbf{m}|-1)}{2}$ and $p(\mathbf{m},\mathbf{m}')$ is the hypergeometric probability mass function. See [3,10]. See also Lemma B.1 in [4] and Sect. 3.3 in [7].

Notice that (7) depends on $\boldsymbol{\alpha}$ only through the rates $\lambda^{\boldsymbol{\alpha}}_{|\mathbf{m}|}$, whose limiting behaviour when $\alpha_1 \to 0$ and $\alpha_2 \to 0$ is given by

$$\lambda^0_{|\mathbf{m}|} := \lim_{\boldsymbol{\alpha}\to(0,0)} \lambda^{\boldsymbol{\alpha}}_{|\mathbf{m}|} = \begin{cases} \binom{|\mathbf{m}|}{2} & \text{if } |\mathbf{m}| > 1 \\ 0 & \text{if } |\mathbf{m}| = 1 \end{cases}, \tag{8}$$

so that the resulting death process has an absorbing state in 1 (cf. [3]). Hence the limiting transition probabilities can be written as

$$p^0_{\mathbf{m},\mathbf{m}'}(t) := \lim_{\boldsymbol{\alpha}\to(0,0)} p^{\boldsymbol{\alpha}}_{\mathbf{m},\mathbf{m}'}(t) = \begin{cases} 0 & \text{if } \mathbf{m}' = (0,0) \\ e^{-\lambda^0_{|\mathbf{m}|}t} & \text{if } \mathbf{m}' = \mathbf{m} \\ C^0_{|\mathbf{m}|,|\mathbf{m}'|}(t)p(\mathbf{m},\mathbf{m}') & \text{if } (0,0) < \mathbf{m}' < \mathbf{m} \end{cases}, \tag{9}$$

where $C^0_{|\mathbf{m}|,|\mathbf{m}'|}(t)$ is the transition probability of a death process with rates $\lambda^0_{|\mathbf{m}|}$ as in (8). These are of course the transition probabilities of Kingman's coalescent without mutation [6]. Thus, unlike in (7), the state $(0,0)$ cannot be reached by the limiting death process, and $(0,1)$ and $(1,0)$ become absorbing states. From a statistical viewpoint, this is intuitive in that the non-informative choice forces at least one of the observations from each of the past and future samples to be used for inference at time t.

2.2 The Aggregation Terms

By Theorem 5 in [4], the term $K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$ in (6) can be written

$$K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}} = \frac{\mu^{\boldsymbol{\alpha}}(\mathbf{m}+\mathbf{n})}{\mu^{\boldsymbol{\alpha}}(\mathbf{n})\mu^{\boldsymbol{\alpha}}(\mathbf{m})}, \quad \mu^{\boldsymbol{\alpha}}(\mathbf{m}) = \frac{\Gamma(\alpha_1+m_1)\Gamma(\alpha_2+m_2)}{\Gamma(\alpha_1+\alpha_2+m_1+m_2)}, \tag{10}$$

with $\mathbf{n} \leq \mathbf{n}_0$ and $\mathbf{m} \leq \mathbf{n}_T$. Here $\mu^{\boldsymbol{\alpha}}(\mathbf{m})$ is the marginal likelihood of collecting a sample with multiplicities $\mathbf{m}$ from the Beta-Bernoulli model in (3). This means that we can interpret $K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$ as the ratio of the likelihood of observing a joint sample with multiplicities $\mathbf{m}+\mathbf{n}$ over the likelihood of observing two samples with multiplicities $\mathbf{m}$ and $\mathbf{n}$ respectively. Thus, a higher $K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$ means that jointly observing $\mathbf{m}+\mathbf{n}$ is more likely, relative to collecting them separately. This interpretation is consistent with the fact that inference on the underlying Wright–Fisher diffusion at time t aims at combining information collected at prior and later times 0 and T.

Letting $\boldsymbol{\alpha} \to (0,0)$ yields a ratio of marginal likelihoods associated to non-informative priors for model (3). As it turns out, we need to analyze separately different scenarios, depending on the values of the observed multiplicities $\mathbf{n}_0$ and $\mathbf{n}_T$. Denote $K^0_{\mathbf{n},\mathbf{m}} := \lim_{\boldsymbol{\alpha}\to(0,0)} K^{\boldsymbol{\alpha}}_{\mathbf{n},\mathbf{m}}$ and $\langle \mathbf{m}, \mathbf{n}\rangle = m_1 n_1 + m_0 n_0$. Then we have the following:

- if $\langle \mathbf{n}_0, \mathbf{n}_T\rangle > 0$ then $K^0_{\mathbf{n},\mathbf{m}} > 0$ if and only if

$$n_{0,i} n_{T,i} > 0 \Rightarrow n_i m_i > 0. \tag{11}$$

  The interpretation is that when there are values observed at both collection times, then $\mathbf{n}, \mathbf{m}$ must feature at least one observation for each shared value, while value not shared can be discarded.
- if $\langle \mathbf{n}_0, \mathbf{n}_T\rangle = 0$ then $K^0_{\mathbf{n},\mathbf{m}} > 0$ if and only if

$$n_{0,i} > 0 \Rightarrow n_i > 0 \text{ and } n_{T,i} > 0 \Rightarrow m_i > 0. \tag{12}$$

  The interpretation here is instead that when no values are shared by the samples at different collection times, at least one observation for each value is retained.

2.3 The Mixture Components

The last term to analyze in (5) is given by the mixture components. From the previous paragraphs we know that the multiplicity vector $(0,0)$ yields a null probability in the non-informative limit. Thus, it makes sense to consider the weak convergence

$$\text{Beta}(\alpha_1 + m_1 + n_1, \alpha_2 + m_0 + n_0) \Rightarrow \text{Beta}(m_1 + n_1, m_0 + n_0),$$

as $\boldsymbol{\alpha} \to (0,0)$, where we identify $\text{Beta}(1,0)$ and $\text{Beta}(0,1)$ as Dirac measures in 1 and 0 respectively.

3 Concluding Remarks

From the above discussion we can draw some conclusion. In a non-informative setting relative to model (1), we are interested in how to perform inference on

the distribution of X_t given vectors of observations $\boldsymbol{Y}_0, \boldsymbol{Y}_T$ collected at times 0 and T with respective multiplicities $\mathbf{n}_0, \mathbf{n}_T$. Our findings are that

$$X_t|\boldsymbol{Y}_0, \boldsymbol{Y}_T, \mathbf{n}, \mathbf{m} \sim \text{Beta}(m_1 + n_1, m_0 + n_0)$$

where the distribution of vector pairs $\mathbf{n}, \mathbf{m}$ is

$$q(\mathbf{n}, \mathbf{m}) \propto p^0_{\mathbf{n}_0,\mathbf{n}}(t) p^0_{\mathbf{n}_T,\mathbf{m}}(T-t) K^0_{\mathbf{n},\mathbf{m}},$$

with $p^0_{\mathbf{n}_0,\mathbf{n}}(t)$ as in (9) and $K^0_{\mathbf{n},\mathbf{m}}$ as in (11) and (12).

A non-informative prior choice has therefore two main effects. It forces inference at the time of interest to depend on at least some of the observed data, retaining at least one instance of each value observed at both times. Exactly how many data points of those observed are retained in the distribution is driven by the mixing probabilities, which are in turn governed by Kingman's coalescent without mutation.

References

1. Ascolani, F., Lijoi, A., Ruggiero, M.: Predictive inference with Fleming-Viot-driven dependent Dirichlet processes. Bayesian Anal. **16**, 371–395 (2021)
2. Ascolani, F., Lijoi, A., Ruggiero, M.: Smoothing distributions for conditional Fleming-Viot and Dawson-Watanabe diffusions. Bernoulli **29**, 1410–1434 (2023)
3. Griffiths, R.C.: Asymptotic line-of-descent distributions. J. Math. Biology **21**, 67–75 (1984)
4. Kon Kam King, G., Papaspiliopoulos, P., Ruggiero, M.: Exact inference for a class of hidden Markov models on general statespaces. Electr. J. Statist. **15**, 2832–2875 (2021)
5. Kon Kam King, G., Pandolfi, A., Piretto, M., Ruggiero, M.: Approximate filtering via discrete dual processes. Stoch. Proc. Appl. **168**, 104268 (2024)
6. Kingman, J.F.C.: The coalescent. Stoch. Proc. Appl. **13**, 235–248 (1982)
7. Papaspiliopoulos, O., Ruggiero, M.: Optimal filtering and the dual process. Bernoulli **20**, 1999–2019 (2014)
8. Papaspiliopoulos, O., Ruggiero, M., Spanò, D.: Conjugacy properties of time-evolving Dirichlet and gamma random measures. Electron. J. Stat. **10**, 3452–3489 (2016)
9. Rubin, D.: The Bayesian Bootstrap. Ann. Stat. **9**, 130–134 (1981)
10. Tavaré, S.: Line-of-descent and genealogical processes, and their applications in population genetic models. Theoret. Population Biol. **26**, 119–164 (1984)

Parents or Peers? Generational Socialisation and Attitudes Towards Migrants Across European Societies

Leo Azzollini[1], Daniela Bellani[2(✉)], and Giulia Rivellini[2]

[1] Department of Sociology and Social Research, University of Trento, Trento, Italy
[2] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milano, Italy
daniela.bellani@unicatt.it

Abstract. The integration of migrants represents still a topical issue of difficult resolution across several countries. Demographic studies have uncovered the roles played by parental socio-economic background and by birth cohort in shaping prejudicial or tolerant attitudes toward immigrants. In this study, we use data from European Social Survey data, Rounds 1–10 (2002–2020). In particular, we make use of the question 'Is the [country] made a worse or a better place to live by people coming to live here from other countries?' to examine the influence of parental socio-economic background on respondents' attitudes towards migrants. Moreover, we study whether this influence varies by birth cohort. Results of linear regression models including country-year fixed effects indicate that, on one side, individuals of recent decades of birth are more pro-immigrants while on the other, the gap in the parental class gradient has widened.

Keywords: Social stratification · birth cohorts · attitudes toward immigrants

1 Motivation

Migration, together with fertility and mortality, is one of the three central dynamics for population studies. Despite numerous empirical evidence on the positive effects of international mobility on the demographic and social wellbeing of the country, the integration of migrants represents still a topical issue of difficult resolution across several countries (AISP 2021). Therefore, understanding how attitudes towards migrants are formed is a central undertaking in contemporary demographic research agenda (Arango et al. 2009; Gereke et al. 2022).

In this study, our goal is to address this issue, examining the roles played by parental socio-economic background and by birth cohort in shaping prejudicial or tolerant attitudes toward immigrants. Considerable research has focused separately on the role played by parents on one hand (Paskov et al. 2021) and by birth cohort on the other (Munck et al. 2018; Jeannet and Drazanova 2023), without examining them together. To the best of our knowledge, our study is the first to disentangle and then combine the roles of inter- (parental socio-economic background background) and intra-generational

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 98–103, 2025.
https://doi.org/10.1007/978-3-031-64431-3_17

(birth cohort) socialization in shaping attitudes towards migrants in a comparative setting. We rely on the interdisciplinary theoretical lens of political demography (Braungart and Braungart 1984; Goerres and Vanhuysse 2021), integrated with theoretical perspectives from research on cohorts (Mannheim 1952; Ryder 1965) to study attitudes towards a phenomenon of crucial relevance for both the population and politics of European countries.

2 Theoretical Framework and Literature Review

The contribution by Braungart and Braungart (1984) on '*Life Course and Generational Politics*' identifies two main ways in which generational factors may affect political attitudes: generational descent and cohort socialization.

The inter-generational perspective highlights parental socio-economic background as the primary driver of offspring's attitudes. This pattern has also been documented for attitudes towards migrants: a lower socio-economic background is typically associated with negative attitudes towards immigrants, that can appear as direct competitors of job positions (Paskov et al. 2021). Therefore:

Hypothesis 1:*respondents with a lower parental socio-economic background are less pro-immigrants than peers with a higher socio-economic background.*

The *birth cohort politics/political generations* perspective envisages instead intra-generational peers as the leading agent of attitude formation. Individuals sharing the same birth years develop similar socio-political attitudes (Elder and George 2016). Applying the classic concept of cohorts as '*engines of social change*' (Ryder 1965) to attitudes towards migration, research has shown that recent cohorts tend to be more pro-immigrants, especially in the context of European integration, with the opening of borders within the European Union (Schmidt 2021). Following this perspective:

Hypothesis 2:*respondents born in more recent birth cohorts are more pro-immigrants than respondents born in older birth cohorts.*

We can reconcile these perspectives thanks to the approach by Mannheim (1952), who argued that birth cohort intersects with other individual social characteristics, such as parental socio-economic background, in shaping attitudes and behavior. By applying this concept to attitudes towards immigrants, we can envisage that while individuals belonging to more recent birth cohorts may still be more pro-migrants on average, their parental characteristics may play an important role in shaping their within-cohort variability. Therefore:

Hypothesis 3:*the influence of parental socio-economic background on attitudes towards migrants, while it persists across older and more recent cohorts, exert heterogeneous effects.*

3 Data and Analytical Strategy

To test our hypotheses, we rely on the entire available European Social Survey, Rounds 1–10 (2002–2020). The ESS is a repeated cross-section fielded biannually to several countries in Europe and its geographical neighborhood (*e.g.*, Turkey). To operationalize attitudes towards immigrants, we rely on the following survey question, from which we

derive our dependent variable: '*Is the [country] made a worse or a better place to live by people coming to live here from other countries?'* (ESS Questionnaire). Respondents are asked to answer on a 11-point scale from 0 (worst) to 10 (best). This question is well-established in the literature as an indicator of attitudes towards immigrants (see Davidov *et al.* 2015). From a total sample size of 463.408 individuals with a valid answer for the dependent variable, we rely on a sample size of 454.259, due to (small proportion of) missing values for certain variables. Respondents are drawn from 38 countries.[1]

Our first key covariate is **decade of birth,** the aggregation of ten yearly birth cohorts**:** *e.g.*, 1950 for someone born between 1950 and 1959. We rely on decade of birth to avoid sample numerosity problems in country-years, and to ensure no cells are empty in the interaction with parental social class. The second key covariate is **parental social class,** with which we operationalize parental socio-economic background according to the EGP class schema,[2] recoded into three large classes (Salariat – 13% of the sample, Intermediate – 32%, Working Class and Missing Social Class[3] –55%). We rely on the highest class of the parents of the respondent. We then include other individual (respondent) socio-demographic controls (Social Class, Level of Education - reduced ISCED - Household Income Decile, Ethnic Minority Status – whether the respondent was born or not in the country of residence – belonging to a religion or not, Gender.)

In terms of **analytical strategy**, we linearly regress attitudes towards immigrants (0–10, worst to best) on the selected covariates and controls, including fixed effects for Country-Years (with clustered standard errors). The use of linear regressions to study attitudes towards immigrants is again well-established in the literature (e.g., Billiet et al., 2014). Formally:

$$Y_{ict} = \beta_0 + \boldsymbol{\gamma'}\, \boldsymbol{X'_{ict}} + \lambda_{ct} + \epsilon_{ict} \tag{1}$$

where Y_{ict} is the dependent variable observed on individual i, for country c, in year t; γ is the vector of coefficients for the vector of $\boldsymbol{X_{ict}}$ sociodemographic variables (which include $\beta_1\, ParentalSocialClass_{\boldsymbol{ict}}$ and $\beta_2\, DecadeOfBirth_{\boldsymbol{ict}}$, our key covariates), and λ_{ct} captures the fixed effects for country c in year t. In a second step we will add in (1) an **interaction term** between $\beta_1\, ParentalSocialClass_{\boldsymbol{ict}}$ and $\beta_2\, DecadeOfBirth_{\boldsymbol{ict}}$, that is $\beta_3\, (ParentalSocialClass\ x\ DecadeOfBirth)_{ict}$.The purpose of the country-year fixed effects is to control for the entire (time-invariant and time-variant) unobserved heterogeneity at the country level, which is a powerful confounder of attitudes towards migrants (e.g., Paskov et al. 2021), allowing us to model the effect of the predictors.

4 Preliminary Results and Conclusions

Figure 1 presents differences in terms of attitudes towards immigrants, considering the regression coefficients for Parental Social Class (on the left) and those for Decade of Birth (on the right). These coefficients are drawn from regression Eq. (1). Two key

[1] Full list available here: https://www.europeansocialsurvey.org/about/participating-countries.

[2] Based on ISCO occupational codes and recoded using the *iscogen* STATA package.

[3] A considerable proportion of respondents (27%) does not report the social class of any of their parents. It is known that these usually represent individuals with the most disadvantaged backgrounds (such as parental unemployment), and therefore best fit with the *Working Class* category in this three-class schema).

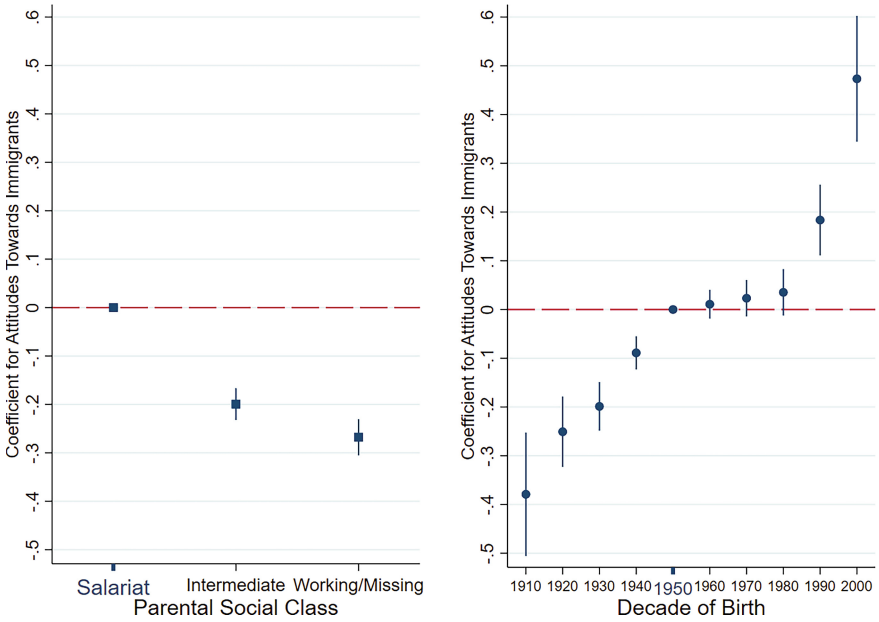


Fig. 1. Coefficients with 95% CIs. Data source: ESS (2002–2020). Design weights included. The categories Salariat and 1950 are the baselines for Parental Social Class and Decade of Birth.

patterns emerge. First, there is a clear parental social class gradient in attitudes towards migrants, with children of parents in the Intermediate and Working classes being less favorable to immigrants than the Salariat class, respectively by −8% and −10% of a Standard Deviation in the dependent variable (which is 2.35). Second, there is significant variation across different Decades of Birth relatively to those born in the 1950s, but this is driven by the groups who were necessarily over 60 (1910–1930) or under 30 (1990–2000) during the study period (2002–2020). We see some less positive coefficients for cohorts 1960–1980, although they are not statistically different from the cohort of the 1950s. At a first glance, the *generational descent perspective* appears to matter more than the cohort perspective in shaping attitudes towards migrants.

Figure 2 presents the Predicted attitudes towards immigrants, obtained with Marginal Effects at the Means after a regression model similar to (1), but with the interaction of the two key covariates. These results highlight how the role of parental social class is consistently strong across different Decades of Birth. While individuals in more recent Decades of Birth are indeed more pro-immigrants, the role of Origin Class is consistent: besides the 1910s and 2000s cohorts, having a Working Class (or missing class) parent is associated with a -0.25 worsening in attitudes towards migrants (around 11% of the SD), relatively to the Salariat within Decade of Birth. Notably, those with an Intermediate Class origin (32% of the sample) shift across Decades from being not statistically different from the Salariat (1930–1950), to being not statistically different from the Working Class, thus widening the class gradient in attitudes towards immigrants in more recent decades of birth.

Both parental background (**H1**) and birth cohort (**H2**) shape attitudes towards immigrants on their own. Their **interactions (H3),** however, highlight that the parental class gradient (the gap in attitudes towards immigrants between the Salariat and the Working Class) *widens* in more recent decades of birth. Yet, the role of birth cohorts is still considerable, considering how those with Working Class origins born in the 1990s have similar attitudes to those with Salariat origins born in the 1950s. Combining the key generational perspectives of generational descent and birth cohort (Braungart and Braungart 1986), we provide interesting insights for the analysis of attitude towards migrants**,** a phenomenon which requires further exploration, considering its rising centrality for the demographic (AISP 2021) and political (Hainmueller and Hopkins 2014) scenarios of European countries.

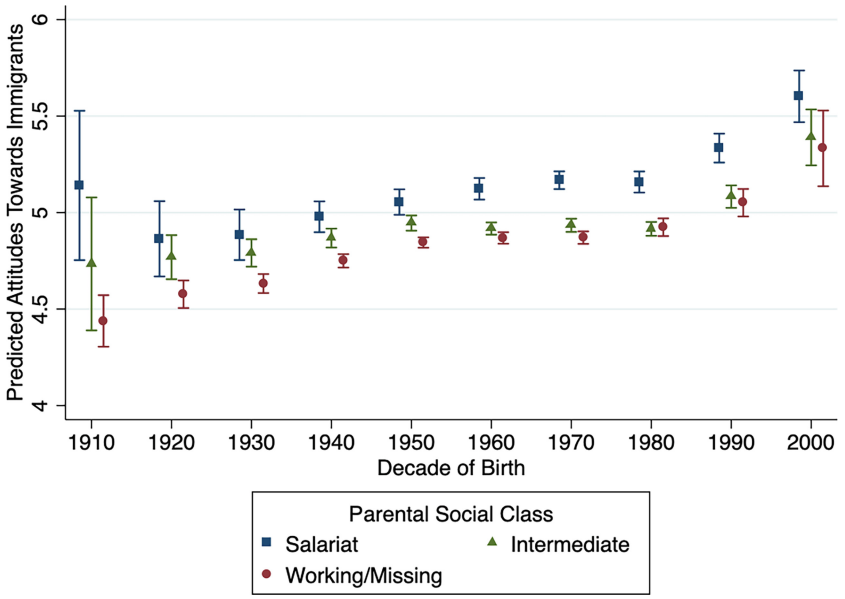


Fig. 2. Predicted Attitudes towards Immigrants with 95% CIs. Joint Impact of Decade of Birth and Origin Social Class. Data source: European Social Survey (2002–2020) (Design weights included).

References

AISP, Rapporto sulla popolazione: L'Italia e le sfide della demografia, il Mulino, Bologna (2021)

Arango, J., et al.: The making of an immigration model: Inflows, impacts and policies in Southern Europe. IDEA working papers, 9 (2009)

Billiet, J., Meuleman, B., De Witte, H.: The relationship between ethnic threat and economic insecurity in times of economic crisis: analysis of European social survey data. Migr. Stud. **2**(2), 135–161 (2014)

Braungart, R.G., Braungart, M.M.: Life-course and generational politics. Ann. Rev. Sociol. **12**(1), 205–231 (1986)

Davidov, E., Seddig, D., Gorodzeisky, A., Raijman, R., Schmidt, P., Semyonov, M.: Direct and indirect predictors of opposition to immigration in Europe: individual values, cultural values, and symbolic threat. J. Ethn. Migr. Stud. **46**(3), 553–573 (2020)

Elder, G.H., Jr., George, L.K.: Age, cohorts, and the life course. In: Handbook of the Life Course, vol. II, pp. 59–85. Springer International Publishing, Cham (2016)

Gereke, J., Schaub, M., Baldassarri, D.: Immigration, integration and cooperation: experimental evidence from a public goods game in Italy. J. Ethn. Migr. Stud. **48**(15), 3761–3788 (2022)

Glenn, N.D.: Distinguishing Age, Period, and Cohort Effects, pp. 465–476. Springer, US (2003)

Goerres, A., Vanhuysse, P.: Global Political Demography: The Politics of Population Change, p. 459. Springer Nature (2021)

Hainmueller, J., Hopkins, D.J.: Public attitudes toward immigration. Annu. Rev. Polit. Sci. **17**, 225–249 (2014)

Jeannet, A.M., Dražanová, L. : Blame it on my youth: the origins of attitudes towards immigration. Acta Politica, 1–30 (2023)

Mannheim, K.: The problem of generations. In: Essays on the Sociology of Knowledge, pp. 276–320. London, Routledge & Kegan Paul (1952)

Munck, I., Barber, C., Torney-Purta, J.: Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: the alignment method applied to IEA CIVED and ICCS. Sociol. Meth. Res. **47**(4), 687–728 (2018)

Paskov, M., Präg, P., Richards, L.: Does downward social mobility make people more hostile towards immigrants? Res. Soc. Stratification Mobility **72**, 100543 (2021)

Ryder, N.B.: The cohort as a concept in the study of social change. Am. Sociol. Rev. **30**, 843–861 (1965)

Schmidt, K.: The dynamics of attitudes toward immigrants: Cohort analyses for Western EU member states. Int. J. Comp. Sociol. **62**(4), 281–310 (2021)



Variable Selection in Multilevel Models via Knockoffs: the Case of INVALSI National Tests

Silvia Bacci, Emanuela Dreassi, Leonardo Grilli(✉), and Carla Rampichini

Department of Statistics, Computer Science, Applications 'G.Parenti', University of Florence, Firenze, Italy
leonardo.grilli@unifi.it

Abstract. We aim to evaluate the performance of the knockoffs method to select the predictors in a model for math achievement in grade 5. We exploit a rich dataset collected by INVALSI with several student background variables. The task is complicated by the multilevel nature of the model, with students nested into schools, and by the missing values in some predictors. To reduce the dependence on Monte Carlo variability, we use derandomized knockoffs. The preliminary results show that, for large-scale assessment data, the proposed approach is feasible and a valuable alternative to traditional model selection methods.

Keywords: education · hierarchical data · latent variable · missing data

1 Introduction

Selecting the covariates in a statistical model is a challenging task. Knockoffs [1] have been proposed to perform the selection while controlling for the overall risk of false negatives, i.e., including irrelevant covariates. This method has been recently adopted in the context of large-scale assessment data [9] to select covariates with missing values. We exploit knockoffs to select the covariates in a model for math achievement at grade 5 using data collected by INVALSI, a section of the National System of Evaluation of the Italian Education and Training System whose main task is to conduct periodic and systematic tests on students' knowledge and skills. These standardized tests measure students' achievement in Italian, Mathematics and English (reading and speaking) [3].

We are interested in detecting the determinants of the ability achieved by pupils in math at the end of primary school (grade 5). The data have a hierarchical structure, with pupils nested into schools and covariates measured at both pupil and school levels. The contribution aims to extend the method of knockoffs for selecting variables to the multilevel model framework.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 104–108, 2025.
https://doi.org/10.1007/978-3-031-64431-3_18

2 Data

We focus on data from the INVALSI sample survey of the year 2022–23 on pupils in grade 5, corresponding to the last year of primary school. The dataset includes 16,828 students nested in 501 schools.

Specifically, we consider the achievement in math, which is a latent ability measured by a set of dichotomously scored items (1: correct answer; 0: wrong answer). The selection of the set of items relies on internationally validated methods based on the Rasch model [4]. For this reason, the ability of a student is measured by the raw score (i.e., the total number of correct answers to the test items). INVALSI provides the response pattern to the items and a Rasch-based estimate of the ability. The data set includes several background variables about the students, such as gender, siblings, parent's educational level and job condition. In addition, a wide set of indicators provides information on student material deprivation, motivation and interest in learning, and relations with classmates. School characteristics include information on the geographical area and the average number of students per class.

A critical issue of the data set is the relevant rate of missing values for many background variables, ranging from 1.6% to 26%. As a preliminary analysis, we consider pupils with complete data, amounting to 8,045 pupils in 436 schools. Figure 1 reports the distribution of the achieved score on mathematics on the selected subsample of pupils with complete data. The distribution is approximately normal, except for some values on the tails. The mean score in math is 198.67 points, with a standard deviation of 39.86.

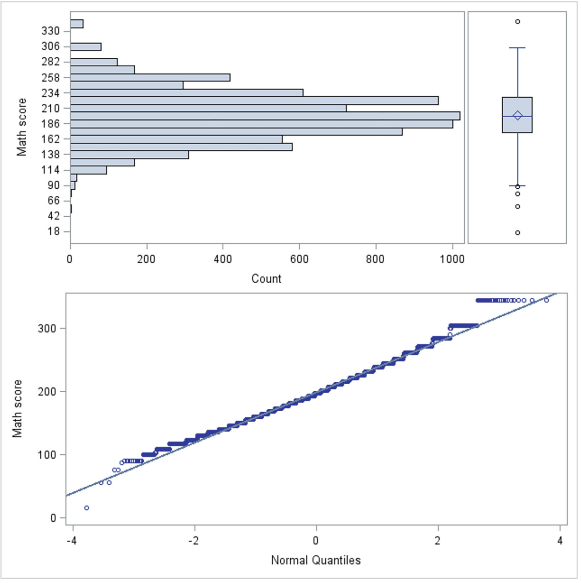


Fig. 1. Math score Distribution

3 Method

The knockoff filter is a general framework for controlling the false discovery rate when performing variable selection [1]. The idea is to generate, for each predictor, an artificial version (knockoff) that is unrelated to the response yet preserving the predictors' correlation structure. A predictor is selected if it is substantially better than its knockoff copy according to some measure of variable importance.

Knockoffs have been recently proposed in the context of large-scale assessment data for selecting the predictors of a structural equation model for student achievement [9]. The novelty in the proposal by Xie et al. [9] lies in extending the knockoffs method to account for missing values in the predictors. To this end, the authors developed an algorithm to generate derandomized knockoffs [6] with missing values under a suitably extended MAR condition.

4 Preliminary Results

We apply the knockoff filter for variable selection to our data using both the Model-X method as implemented in the R package `knockoff` [7] and the derandomized knockoffs procedure implemented in the R package `derandomKnock` [5].

Table 1 reports the t-values of a linear regression model for the math score fitted on the covariates selected by different methods, namely Model-X knockoffs with $FDR = 0.10$, derandomized knockoffs with $PFER = 1$ and threshold for selection 0.5, and a standard stepwise procedure based on BIC. All the covariates are binary except for the count variables denoted with #. The three methods lead to the selection of substantially the same covariates. In particular, the derandomized knockoff procedure reduces the number of predictors from 43 to 25.

In line with other surveys, males have better results compared to females. Also, pupils who speak Italian at home and with friends obtain higher scores in math. Moreover, home environment, parents' education, and parents' occupational status are relevant determinants of pupils' ability.

These preliminary results are obtained on the subset of complete observations, that is, on 8,045 out of the 16,828 pupils. The next step in our analysis is to deal with missing values using multiple imputation [8], comparing it with the approach proposed by Xie et al. in [9]. In addition, since pupils are nested within schools, their responses are correlated, and we can exploit information on teachers and school characteristics to identify determinants of math ability at the school and pupil levels. Thus, we aim to extend the derandomized method for variable selection in a two-level model with missing values on covariates [2].

Table 1. Variable selection via Knockoffs and Stepwise methods: t-values associated to the selected covariates ($n = 8045$)

	Knockoffs		Stepwise
Selected covariates	Model-X	Derandom.	
Male	11.78	11.71	11.65
Regular student	2.87	2.78	
Nursery school	5.18	5.09	5.45
Italian language			
used at home	3.51	3.37	4.01
used with friends	4.66	4.62	4.74
At home, you have:			
# brothers	–4.62	–4.68	–4,61
# sisters		–1.95	
# of books	13.40	13.63	14.36
computer	3.80	3.99	4.60
software	1.55		
internet connection	2.60	2.61	
personal room	–1.68		
dictionary	4.99	4.99	4.99
tablet	–4.69	–4.90	–4.59
smartphone	–5.70	–5.79	–6.25
Mother's education			
high school	4.86	4.76	4.78
graduate	5.50	5.42	6.79
Father's education			
high school	4.64	4.52	4.75
graduate	7.02	6.99	7.21
Mother's occupation			
manager	2.98	2.89	
white collar	2.43	2.30	
Father's occupation			
unemployed	–3.43	–3.35	–3.83
blue collar	–2.68	–2.68	–3.56
School area			
center	5.21	5.24	6.33
sud	6.19	6.26	7.44
islands	–2.27	–2.22	
# selected covariates (out of 43)	25	24	18

References

1. Candès E., Fan Y., Janson L., Lv J.: Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J. Roy. Stat. Soc. Series B (Stat. Meth.) **80**(3), 551–577 (2018). https://DOI.org/10.1111/rssb.12265
2. Grilli, L., Marino, F.M., Paccagnella, O., Rampichini, C.: Multiple imputation and selection of ordinal level 2 predictors in multilevel models: An analysis of the relationship between student ratings and teacher practices and attitudes. Stat. Model. **22**(3), 221–238 (2022). https://doi.org/10.1177/1471082X20949710
3. Cardone, M., Falzetti, P., Sacco, C.: INVALSI data for school system improvement: the value added. Working Papers INVALSI, n. 5694, Invalsi (2019)
4. Rasch, G.: Probabilistic Models for some Intelligence and Attainment Tests, Danish Institute for Educational Research, Copenhagen (1960)
5. Ren, Z.: DerandomKnock: an R package to implement the derandomized knockoffs procedure. R package version 0.1.0 (2022). https://github.com/zhimeir/derandomKnock. https://zhimeir.github.io/derandomKnock/index.html
6. Ren Z., Wei Y., Candès E.: Derandomizing knockoffs. J. American Stat. Assoc. **118**(542), 948–958 (2023). https://DOI.org/10.1080/01621459.2021.1962720
7. Sesia, M.: The Knockoff Filter for Controlled Variable Selection. R package Version 0.3.6 (2022). https://cran.r-project.org/web/packages/knockoff/index.html
8. van Buuren, S.: Flexible Imputation of Missing Data, 2nd edn. CRC Press Taylor & Francis Group, Boca Raton, FL (2018)
9. Xie, Z., Chen, Y., von Davier, M., Weng, H.: Variable selection in latent variable models via knockoffs: an application to international large-scale assessment in education. J. Roy. Stat. Soc. Series A Stat. Soc. (2023). DOIurl10.1093/jrsssa/qnad137



A Fully Rank-Based Hierarchical Clustering Approach Using Information Imbalance

Elena Ballante[1(✉)], Antonietta Mira[2,3], and Silvia Figini[1]

[1] Department of Political and Social Sciences, University of Pavia, Pavia, Italy
elena.ballante@unipv.it
[2] Faculty of Economics, Euler Institute, Università della Svizzera italiana, Lugano, Switzerland
[3] Department of Science and High Technology, Insubria University, Como, Italy

Abstract. A new clustering algorithm is proposed. It deploys the structure of hierarchical clustering by embedding a new similarity measure between observations. The proposed similarity measure relies on a completely non-parametric approach that involves a rank transformation of the distances.

The preliminary results on synthetic data confirm the validity of the proposed method in terms of its ability to capture the predefined cluster structures in the data.

The advantages of the proposed algorithm are manifold. It offers a clustering approach robust with respect to outliers and specific pattern in the data due to the rank-based approach. A high level of flexibility is involved, since the rank transformation can be adapted to every distance or ordering method. Moreover, a natural way to perform variable selection is included.

Keywords: Hierarchical clustering · Ranking · Information imbalance

1 Introduction

Cluster analysis aims to discover internal patterns within a dataset by identifying structures in the form of groups of similar data items. This type of problems arises in numerous applications when the goal is to uncover similarities and differences in unlabeled datasets (*unsupervised learning*). It is a research topic that has been explored over several decades. For example, the k-means algorithm was proposed by Forgy in 1965 [1], and a vast variety of clustering methods already exists. However, the difficulties in defining the *similarity* between subjects, and various nuances this concept can encompass, continue to make it a rich field of research even today. In this work, hierarchical clustering methods will be explored, coupled with a new rank-based method for evaluating subjects' similarity.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 109–114, 2025.
https://doi.org/10.1007/978-3-031-64431-3_19

The rest of the paper is organized as follows. A brief review of the background concepts (hierarchical clustering methods and information imbalance) is provided in Sect. 2, the proposed method is described in Sect. 3, and in Sect. 4 the preliminary results on simulated data are shown. Conclusions are illustrated in Sect. 5.

2 Background

2.1 Hierarchical Clustering

Hierarchical clustering methods rely on the idea of defining clusters based on a dissimilarity/similarity measure between subjects [2].

It is widely used thanks to the flexibility of the method, which can lead to extremely different results depending on the chosen dissimilarity/similarity measure. The most common choice is the Minkowski distance, which comprises Euclidean distance (L^2 norm) and Manhattan distance (L^1 norm).

Moreover the method is considered the optimal choice when the number of clusters is not clearly fixed, since it does not require a preliminary decision.

The main features that characterize a hierarchical clustering model are the direction of the hierarchy (bottom-up/top-down approaches) and the linkage strategy (the more common ones are complete, average and single linkage). The bottom up approach starts from the single observations and couples them with another observation based on the similarity matrix, creating clusters of two similar observations. Then, the most similar couples are grouped, creating bigger clusters, and so on until all the observations are aggregated. The linkage function defines how to measure the distance between clusters. Complete linkage defines the cluster distance as the distance between the most distant observations in the two clusters, single linkage takes the distance between the least distant observations in the two clusters, and average linkage takes the average of the distances of all the observations in the two different clusters.

The hierarchical clustering does not produce a single grouping of the data, but a dendrogram. The dendrogram is a diagram that illustrates the arrangement of data points or clusters in a hierarchical tree-like structure. It typically consists of a vertical axis representing the level of similarity or dissimilarity between clusters or data points, and a horizontal axis representing the individual data points or clusters being merged. The dendrogram can be cut at different heights to produce different numbers of clusters. The appropriate number of clusters can be determined based on domain knowledge, visual inspection of the dendrogram, or using methods like elbow [3] or silhouette score [4]. The structure of clusters is determined by analyzing the dendrogram and identifying the clusters formed at different levels of aggregation.

2.2 Information Imbalance

In addition to the classical hierarchical clustering approach, the concept of *Information Imbalance* (II) is deployed. This concept was introduced in [5] as a

method for variable selection, proposing a measure to explore the relative information content of one distance measure with respect to another. The definition is based of the idea of quantifying the relative quality of two distance measures by analyzing the ranks of the first neighbours of each point. The underlying idea is to assess the relationship between any two distance measures d_A and d_B by using the properties of the conditional rank distribution $p(r^B|r^A = 1)$. The closer this distribution is to a delta function peaked at 1, the more information about space B is contained within space A. The formal definition, deploying copula variables, is stated and described in [5]. The II ranges in $(0, 1)$.

The computational approach leads to a slightly different definition. The II on a dataset with N data points is defined as:

$$\Delta(A \to B) = 2\langle r^B|r^A = 1\rangle/N \tag{1}$$

where $\langle r^B|r^A = 1\rangle$ represents the average rank in space B of the first neighbours in space A of all the variable in the data.

3 Proposed Method: A New Rank-Based Similarity Measure

The hierarchical clustering results are strongly determined by the similarity (or dissimilarity) measure adopted, as the choice of the measure determines both the shape of the dendrogram and the nature of the clusters obtained. For these reasons, the definition of the similarity measure is the focus of this work.

The more common distances adopted in practice are related to Euclidean distance, or its generalization like Minkowsky distance [6]. The idea is to propose a complete non-parametric approach based on ranking. This approach overcomes some limitations of the classical distances adopted, for example it is robust with respect to the presence of outliers and improves the stability of the results.

Moreover, we want to introduce a weighting mechanism to improve the reliability of results. This weighting mechanism deploys the concept of II, described in Sect. 2. The general idea is to assign higher weights to the variables that are more informative with respect to the distance induced by the whole set of variables. The proposed similarity metric is defined as follows:

$$sim(i,j) = \sum_{m=1}^{p}(1 - \frac{\Delta(d_{X_m} \longrightarrow d_{X_1 \ldots X_p})}{\sum_{s=1}^{p}\Delta(d_{X_s} \longrightarrow d_{X_1 \ldots X_p})})(2N - r_{i,j}^m - r_{j,i}^m) \tag{2}$$

where $X_1, ..., X_p$ are the p variables in the dataset, N is the number of observations, and $i, j \in \{1, ..., N\}$ index all pairs of observations in the dataset. $r_{i,j}^m$ represents the rank of the j^{th} observation relative to the i^{th} observation, measured only relative to the m^{th} variable. Each ranking is weighted with respect to the II of the m^{th} variable relative to all the other variables. Weighting the difference in ranks by the II provides an intuitive measure of variable importance. The proposed similarity measure is presented and explored in details in [7].

The similarity measure described can be used to evaluate the similarity matrix for the data at hand, representing the pairwise similarities between all data points in the dataset. Hierarchical clustering with a bottom up approach proceeds by iteratively merging data points or clusters based on their similarities. Initially, each data point is treated as a singleton cluster. Then, at each iteration, the two clusters or data points with the highest similarity are merged into a single cluster. As clusters are merged, a dendrogram is constructed to visualize the hierarchical structure of the clustering process. The dendrogram represents the fusion of clusters at each step and provides insights into the relationships between clusters and their constituents. The dendrogram can be cut at different heights to produce different numbers of clusters.

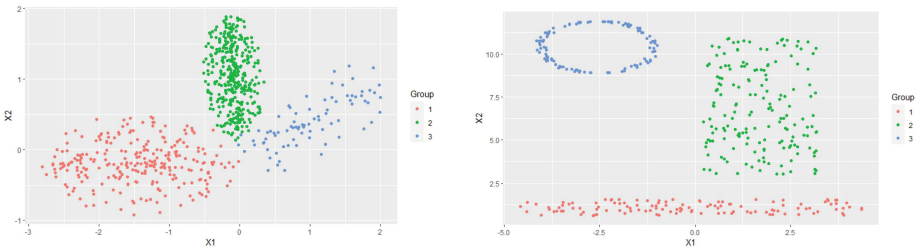
4 Preliminary Results on Synthetic Data

In order to show the specific performances of the proposed method in comparison to the most common clustering algorithms, we present two case studies with different characteristics. The datasets are provided by Parmar in the GitHub repository milaan9/Clustering-Datasets.

The proposed method is compared with the hierarchical clustering using the Euclidean distance and with a classical centroid-based clustering method, the k-means algorithm [1]. Both hierarchical algorithms utilize a bottom-up approach with complete linkage.

Both datasets considered have a target label, so the performance of the algorithms will be evaluated based on their ability to reproduce the ground-truth label. We are not considering the choice of the optimal number of clusters, as this represents a future development of the method. In the case studies, the number of groups is treated as a known parameter.

The datasets are depicted in Fig. 1.



(a) First case study (b) Second case study

Fig. 1. Graphical representation of the datasets considered to assess the proposed method.

4.1 First Case Study

The dataset for the first case study, with the true labels, is shown in Fig. 1a. It consists of a synthetic dataset composed by $N = 715$ observations, two quantitative variables, and the ground-truth label. The groups are unbalanced: 264 observations (36.9%) are in group 1, 370 (51.7%) in group 2, and only 81 (11.3%) in group 3.

The results of the clustering methods are described in Table 1, where we show the confusion matrices with the true labels on the rows and the predicted labels on the columns. It is easy to observe that no method is optimal in this case. However, the proposed method achieves the best results, perfectly identifying labels in groups 1 and 3, with only 35 errors regarding group 2. In contrast, the classical hierarchical approach with Euclidean distance performs worse than k-means and the proposed method. It is completely unable to identify labels in group 3, and also struggles in the identification of labels in group 1. k-means algorithm shows intermediate performance.

Table 1. First case study: comparison of the three clustering approaches by confusion matrix where columns are indexed by predicted labels while rows are indexed by the true ones.

	1	2	3
1	257	7	0
2	0	370	0
3	0	28	53

(a) Hierarchical clustering with proposed similarity

	1	2	3
1	167	1	96
2	0	370	0
3	0	81	0

(b) Hierarchical clustering with Euclidean distance

	1	2	3
1	251	0	13
2	0	288	82
3	0	0	81

(c) k-means algorithm

4.2 Second Case Study

The second case study is shown in Fig. 1b. It is a synthetic dataset composed by $N = 400$ observations, two quantitative variables, and the ground-truth label. The groups are almost balanced: 120 observations (30.0%) are in group 1, 170 (42.5%) in group 2 and 110 (27.5%) in group 3.

The results of the clustering methods are described in Table 2, where we display the confusion matrices with the true labels on the rows and the predicted labels on the columns. The hierarchical approach with Euclidean distance is the most problematic. It correctly identifies most labels in group 3 but makes errors regarding both group 1 and 2. The proposed algorithm shows similar performances compared to k-means, making 28 and 27 mistakes, respectively. Both of them show difficulties in identifying labels in group 2.

Table 2. Second case study: comparison of the three clustering approaches by confusion matrix where columns are indexed by predicted labels while rows are indexed by the true ones.

	1	2	3
1	120	0	0
2	28	142	0
3	0	0	110

(a) Hierarchical clustering with proposed similarity

	1	2	3
1	44	76	0
2	0	89	81
3	0	0	110

(b) Hierarchical clustering with Euclidean distance

	1	2	3
1	120	0	0
2	23	143	4
3	0	0	110

(c) k-means algorithm

5 Conclusion

In this work, a new approach to hierarchical clustering is proposed. We introduce a novel measure of similarity between statistical units, leveraging the main advantages of a completely rank-based non-parametric approach embedded in a hierarchical algorithm.

The preliminary results confirm the validity of the proposed method, but further investigations are needed. For example, the influence of the clustering strategy (top-down or bottom-up) and linkage methods should be explored. Furthermore, the flexibility of the method is not fully explored, as the ranking approach was developed on the basis of Euclidean distance, but it can be generalized to any distance or ordering method. Additionally, the influence of II weights warrants further exploration, and it can also be deployed as a variable selection approach. Another area of future research is the development of a method for estimating the optimal number of clusters. The R code will be available soon in a GitHub repository.

References

1. Forgy, E.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics **21**, 768–769 (1965)
2. Jain, A.K., Dubes, R.C.: Algorithms for clustering data, Prentice-Hall, Inc. (1988)
3. Joshi, K.D., Nalwade, P.S.: Modified K-means for better initial cluster centres. Int. J. Comput. Sci. Mob. Comput. **2**, 219–223 (2013)
4. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
5. Glielmo, A., Zeni, C., Cheng, B., Csanyi, G., Laio, A.: Ranking the information content of distance measures, PNAS Nexus, **1**(2), pgac039 (2022)
6. Nielsen, F.: Introduction to HPC with MPI for Data Science. UTCS, Springer, Cham (2016). https://doi.org/10.1007/978-3-319-21903-5_9
7. Ballante, E., Figini, S., Mira, A.: Non-parametric similarity measure based on information imbalance. Technical Report (2024)



Non-parametric Similarity Measure Based on Information Imbalance

Elena Ballante[1(✉)], Silvia Figini[1], and Antonietta Mira[2,3]

[1] Department of Political and Social Sciences, University of Pavia, Pavia, Italy
elena.ballante@unipv.it
[2] Faculty of Economics, Euler Institute, Università della Svizzera italiana, Lugano, Switzerland
[3] Department of Science and High Technology, Insubria University, Varese, Italy

Abstract. Measuring the similarity between objects is an important task in many different statistical and machine learning problems. In this work, we propose a new similarity measure. The proposal is based on the idea of rank transformation of distances and deploys the concept of Information Imbalance as a measure of variable importance.

The specific characteristics captured by the proposed similarity are demonstrated in a toy example, where it is compared with the Euclidean distance and the cosine similarity on a two dimensional dataset.

The proposed similarity can be embedded in clustering algorithms, k-Nearest Neighbour type models, or it can be tested in other applications such as image analysis.

Keywords: Similarity measure · Ranking · Information Imbalance

1 Introduction

The measurement of similarity between objects is crucial in various statistical and machine learning problems, such as image analysis, pattern recognition, classification, clustering, and information retrieval. For example, it constitutes a fundamental operation in models like k-Nearest Neighbour, hierarchical clustering, k-means, and others.

Although a unique definition does not exist, we can generally consider a similarity measure as a real-valued function that quantifies the similarity between two objects. While not necessary for its definition, a similarity function can be the inverse of distance metrics.

Numerous similarity measures have been proposed and researched in literature, because it is not feasible to provide a single definition of similarity for every application context. Various comparative studies are presented in the literature to assist researchers in navigating this topic, such as [1,2].

In this work, we propose a new similarity measure. The proposal is based on the idea of rank transformation of distances and deploys the concept of Information Imbalance as a measure of variable importance.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 115–120, 2025.
https://doi.org/10.1007/978-3-031-64431-3_20

The rest of the paper is organized as follows. The background concept about Information Imbalance is provided in Sect. 2, the proposed method is described in Sect. 3, and a toy example is presented in Sect. 4. Conclusions and further ideas of research are discussed in Sect. 5.

2 Background: Information Imbalance

The concept of *Information Imbalance* (II) is deployed for the definition of the similarity measure. This concept was introduced in [3] as a method for variable selection, proposing a measure to explore the relative information content of one distance measure with respect to another.

The definition is based of the idea of quantifying the relative quality of two distance measures by investigating the ranks of the first neighbours of each point. The underlying idea is to assess the relationship between any two distance measures d_A and d_B by using the properties of the conditional rank distribution $p(r^B|r^A = 1)$. The closer this distribution is to a delta function peaked at 1, the more information about space B is contained within space A. The formal definition is based on copula variables, where we can define a copula variable c_A as the cumulative distribution $C_A = \int_0^{d_A} p_A(w|x)dw$ where $p_A(w|x)$ is the probability of sampling a data point within distance w from x in the A space. Copula variables and distance ranks can be considered continuous-discrete analogues of each other. As a consequence the conditional distribution $p(r^B|r^A = 1)$ is the estimates of $p(c_B|c_A)$ with c_A conditioned to be very small.

Considering this setting, the Information Imbalance from space A to space B is defined as:

$$\Delta(A \to B) = 2 \lim_{\epsilon \to 0} \langle c_B | c_A = \epsilon \rangle$$

where the conditional expectation $\langle c_B|c_A = \epsilon\rangle = \int c_B p(c_B|c_A = \epsilon)d_{c_B}$ is used to characterize the deviation of $p(c_B|c_A = \epsilon)$ from a delta function.

In the computational setting the II on a dataset with N data points is defined as:

$$\Delta(A \to B) \approx 2\langle r^B | r^A = 1\rangle / N \tag{1}$$

where $\langle r^B|r^A = 1\rangle$ is the average rank in space B of the first neighbours with respect to the space A of all the variables in the data.

More theoretical details and examples are provided in [3].

3 A New Rank-Based Similarity Measure

In this section a new similarity measure is proposed on the basis of a complete non-parametric approach that involves rank transformation of distances. Our idea overcomes some limitations of the classical distances adopted, for example it is robust with respect to the presence of outliers.

A weighting mechanism to improve the reliability of results is adopted. This weighting mechanism deploys the concept of II, described in Sect. 2. The general

idea is to assign higher weights to the variables that are more informative with respect to the distance induced by the whole set of variables, deploying the definition of Information Imbalance in this unsupervised context. The contribution of each variable in the data is weighted with respect to the II of the m^{th} variable relative to all the other variables, so that the weights sum to 1. The weights are denoted as $1 - II_{X_m}$, where II_{X_m} is defined as follows:

$$II_{X_m} = \frac{\Delta(X_m \to (X_1, ..., X_p))}{\sum_{j=1}^{p} \Delta(X_j \to (X_1, ..., X_p))} \tag{2}$$

where $\Delta(X_m \to (X_1, ..., X_p))$ corresponds to the definition of Information Imbalance in Eq. 1.

Incorporating the II in the weighting mechanism provides an intuitive measure of variable importance and can also be a crucial point to incorporate a variable selection process.

The proposed similarity metric is as follows:

$$sim(i, j) = \sum_{m=1}^{p} (1 - II_{X_m})(2N - r_{i,j}^m - r_{j,i}^m) \tag{3}$$

where $X_1, ..., X_p$ are the p variables in the dataset, n is the number of observations, and $i, j \in \{1, ..., N\}$ index all pairs of observations in the dataset and $r_{i,j}^m$ represents the rank of the j^{th} observation relative to the i^{th} observation, measured only relative to the m^{th} variable.

The similarity measure is symmetric by definition, i.e. $sim(i, j) = sim(j, i)$, $\forall i, j$.

It is not defined for $i = j$, but we can adopt as convention that $r_{i,i}^m = 0$. In this case the similarity is maximal and takes value $sim(i, i) = 2Np$.

If the observations i and j are the most dissimilar relative to all the other variables $m \in \{1, ...p\}$, then $r_{i,j}^m = r_{j,i}^m = N-1$ and $sim(i, j) = 2p$. If observations i and j are the most similar, they will be mutual first neighbours with respect all the variables $m \in \{1, ...p\}$, so $r_{i,j}^m = r_{j,i}^m = 1$ and $sim(i, j) = p * (2N - 2)$. So the proposed similarity ranges between $2p$ and $2p(N - 1)$ in standard applications, when $sim(i, i)$ is not evaluated.

Clearly a normalized version, ranging between zero and 1, of the proposed similarity can be defined as

$$sim_{norm}(i, j) = \frac{1}{2p(N - 1)} \sum_{m=1}^{p} (1 - II_{X_m})(2N - 2 - r_{i,j}^m - r_{j,i}^m). \tag{4}$$

4 Toy Example: Similarity Evaluation

In order to clarify the specific functioning of the method, a toy example is described. The similarity measure that is evaluated in this section is the normalized definition of Eq. 4.

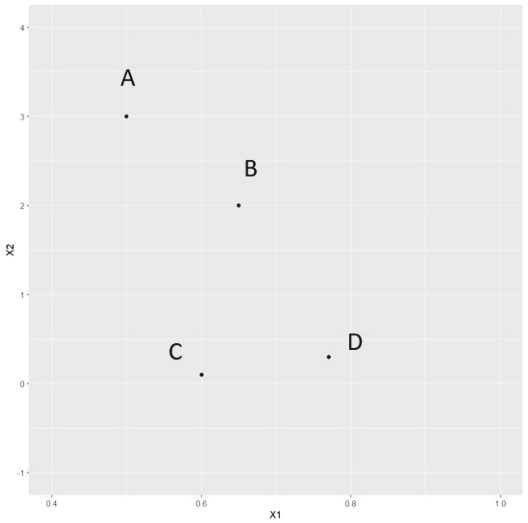


Fig. 1. Graphical representation of the data considered in the toy example.

Consider a dataset of $N = 4$ observations and two features X_1 and X_2 depicted in Fig. 1 and having following coordinates: A(0.5, 3), B(0.65, 2), C(0.6, 0.1), D(0.77, 0.3).

The Information Imbalances evaluated for X_1 and X_2 with respect to the global distance defined by (X_1, X_2) are $\Delta(X_1 \rightarrow (X_1, X_2)) = 1.25$ and $\Delta(X_2 \rightarrow (X_1, X_2)) = 0.5$.

These values of the II mean that variable X_2 is more informative than variable X_1, relative to the Euclidean distances induced by (X_1, X_2). The definition of the normalized weighs is provided in Eq. 2.

A couple of examples of the similarity calculations are the following:

$$\begin{aligned} sim(A,B) &= (1 - II_{X_1}) * (2N - r^1_{AB} + r^1_{BA}) + (1 - II_{X_2}) * (2N - r^2_{AB} + r^2_{BA}) \\ &= (1 - -0.71) * (8 - -2 - 3) + (1 - -0.29) * (8 - -1 - 1) \\ &= 0.29 * 3 + 0.71 * 6 = 5.13 \end{aligned}$$

$$\begin{aligned} sim(A,C) &= (1 - II_{X_1}) * (2N - r^1_{CA} - r^1_{AC}) + (1 - II_{X_2}) * (2N - r^2_{CA} - r^2_{AC}) \\ &= (1 - -0.71) * (8 - -1 - 2) + (1 - -0.29) * (8 - -3 - 3) \\ &= 0.29 * 5 + 0.71 * 2 = 2.87 \end{aligned}$$

$$\begin{aligned} sim(C,D) &= (1 - II_{X_1}) * (2N - r^1_{CD} + r^1_{DC}) + (1 - II_{X2}) * (2N - r^2_{CD} + r^2_{DC}) \\ &= (1 - -0.71) * (8 - -3 - 2) + (1 - -0.29) * (8 - -1 - 1) \\ &= 0.29 * 3 + 0.71 * 6 = 5.13 \end{aligned}$$

The complete evaluation with the normalization and the comparison with the Euclidean distance and cosine similarity is described in Table 1.

Table 1. Comparison of the proposed similarity measure with the most common similarity/distances in literature.

	Euclidean distance	Proposed measure	Cosine similarity
(A,B)	1.01	0.39	0.99
(A,C)	2.90	0.11	0.32
(A,D)	2.71	0.09	0.51
(B,C)	1.90	0.23	0.46
(B,D)	1.70	0.29	0.63
(C,D)	0.26	0.39	0.98

We can observe that the distance/similarity between the pairs (A,C) and (A,D) are similar with respect to both Euclidean distance and the proposed similarity. However, cosine similarity shows an higher value for (A,D) (0.52 vs 0.32). The main difference with respect to the Euclidean distance is shown in the evaluation of difference/similarity between the pairs (A,B) and (C,D). Euclidean distance detects a higher distance in the first pair than in the second (1.01 vs 0.26), whereas the proposed similarity shows the same exact value of 0.39, similar to the cosine similarity.

The cosine similarity shows a difference between the pairs (B,C) and (B,D) (0.46 vs 0.63) that is not detected by the proposed measure (0.23 vs 0.29) and to a lesser degree by the Euclidean distance (1.70 vs 1.90).

It is evident that the three measures capture different aspects in the evaluation of distance/similarity, thereby making them useful in different applications. For instance, while Euclidean distance may excel in scenarios where the spatial relationships between data points are crucial, cosine similarity might be more adept at handling cases involving high-dimensional data or text analysis. Similarly, the proposed similarity measure could offer advantages in contexts where rank-based comparisons or considerations of variable importance are paramount.

5 Conclusion

In this work, we introduce a new measure of similarity between statistical units, leveraging the main advantages of a completely rank-based non-parametric approach and the concept of Information Imbalance. A toy example is provided to demonstrate how the measure works in a simple and controlled setting.

Further investigation are needed; a comparative study involving other classical similarity measures can help understanding the peculiarities of the method. Moreover, a sensitivity analysis on the influence of the dimensionality should be

performed to understand how and when Information Imbalance can be deployed as a variable selection method inside the similarity.

As further avenue of research, the proposed similarity measure can be embedded in clustering and distance based models. A first attempt of application is described in [4].

As a possible extension of this work, the measure can be tested in image comparison.

References

1. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y.: A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS ONE **10**(12), e0144059 (2015)
2. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the eighth SIAM International Conference on Data Mining, pp. 243–254 (2008)
3. Glielmo, A., Zeni, C., Cheng, B., Csanyi, G., Laio, A.: Ranking the information content of distance measures. PNAS Nexus, **1**(2), pgac039 (2022)
4. Ballante, E., Mira, A., Figini, S.: A fully rank-based hierarchical clustering approach using information imbalance. Technical Report (2024)

Data Transformation and Its Validity in a Two-Sample Problem: An Illustration Based on Graphical Models

Erika Banzato[1(✉)], Davide Risso[1], Monica Chiogna[2], and Vera Djordjilović[3]

[1] Department of Statistical Sciences, University of Padua, via C. Battisti 241, Padua, Italy
erika.banzato@unipd.it

[2] Department of Statistical Sciences, University of Bologna, Via Belle Arti, 41, Bologna, Italy

[3] Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, Venice, Italy

Abstract. The rapid expansion of technology in medical and biological domains has led to a surge in available data, accompanied by increased complexity. Classical statistical methods, primarily developed for analyzing data assumed to follow a normal distribution, may prove inadequate for handling this complexity. We focus our attention on transcriptomic data, in particular single-cell RNAseq data, which measure gene expression as counts rather than intensities. Through a simulation study, this study aims to demonstrate the limitations of common data transformations intended to fit data into a normal distribution, highlighting potential misinterpretations. Simulation results show that the transformation of the data might affect the results and their interpretation. In particular, we show that in a two-sample problem in a graphical framework, the set of nodes that differs in the two conditions is affected by the transformations. This behavior might be due to the fact that transformations can distort graphical structures reliant on conditional dependencies among variables, affecting final conclusions. Specifically, when analyzing RNA-seq data, methods tailored for count data are preferable over those designed for normally distributed data. The findings underscore the need for specialized approaches in handling count data in two-sample problems and advocate for further research into alternative methods for differential network analysis.

Keywords: two-sample problem · data transformation · graphical models · count data

1 Introduction

Data transformation represents a controversial tool for statistical analysis. It is often used as a pre-processing step, but little can be found in the literature about

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 121–126, 2025.
https://doi.org/10.1007/978-3-031-64431-3_21

the risks of improper and naive use. This work aims to point out the increasing need for *ad hoc* methods to deal with the new complex types of data now available in the biological and medical fields. New tools should reflect the complex nature of data, instead of trying to transform data to match the assumptions behind the classical statistical methodologies, usually developed for normally distributed data.

Originally, when transcriptomics data started to become available, the technology of choice for measuring gene expression was the microarray assay. This technology optically scans fluorophore intensities, providing data on a continuous scale [1]. This kind of data is commonly assumed to be normally distributed after being transformed in the logarithmic scale, and many statistical techniques have been developed to deal with such data. Later, a new technology allowed for the high-throughput sequencing of RNA molecules (i.e., RNA-seq), and now represents the reference technology for the study of genome-wide transcription levels [2]. One of the main advantages of RNA-seq over microarrays is that it allows the study of gene expression even at single-cell resolution [3]. However, as measures of gene expression, RNA-seq yields counts rather than intensities, which typically show skewed distributions with a high variance. Moreover, they very often show a large number of zero counts, typically larger than expected in a Poisson model [4].

Network analysis is one of the most prominent fields of research, with a notable emphasis on uncovering differences within networks in multiple conditions. The aim is to identify potential differences in the network by comparing data arising from different conditions [5]. However, statistical techniques have been mainly developed for data assumed to be normally distributed, but there is little information about how this applies to count data, such as RNA-seq.

When dealing with such data, it is common to assume normality on the logarithmic scale or on other transformations [6]. Some transformations show good performance when the mean and variance of the variables are relatively high and the presence of ties does not have a strong impact on the analysis. However, the impact of these transformations on more skewed data is not clear. In fact, the main issue with single-cell data is the overabundance of zeros that leads to a high presence of ties, which is a big hurdle when trying to normalize the data. Moreover, in this particular setting, an important aspect is that the adopted transformation should preserve the conditional independence of the variables and potential differences. The primary interest is the characterization of a transformation that preserves the relations between variables while maintaining possible changes in the relationships.

Through a simulation study, this work aims to compare different transformation approaches that are used in practical analyses, to highlight the main advantages and drawbacks when data are transformed. The purpose is to emphasize how the data transformation process could change the conditional independence relationships between variables and lead to misleading results.

2 Simulation Study

We ran a simulation study to compare the performances of different data transformations applied to data from a Poisson graphical model. Let $G = (V, E)$ be the underlying undirected graph, where V is the set of nodes and $E = \{(u, v) : u, v \in V\}$ is the set of edges. A graph is said to be undirected if two vertices a and b are connected by an edge (a, b) in addition to (b, a).

Single-cell RNAseq data can be described using a truncated Poisson graphical model (TPGM) [7,8]. The joint distribution of the p-dimensional random vector $\boldsymbol{Y}$ with domain $Y_i \in \{0, 1, \ldots, D_i\}$ for $i = 1, \ldots, p$ can be written as

$$p(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) \propto \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{\phi} \boldsymbol{y} + \sum_{j=1}^{p} \log(y_j!) \right\} \tag{1}$$

where the matrix ϕ represents the graphical structure of the model.

For illustrative purposes, we consider a simple graph with four nodes and four edges. Data were simulated from the graph in Fig. 1, using the code in [8]. We assumed that the samples come from two TPGM distributions in (1), differing only for the parameter θ_3. Then four transformations were applied to the generated data: logarithm, square root, deviance residuals, and randomized quantile residuals [6]. To avoid computational issues caused by zero counts generated in the logarithmic transformation, a unit was added to each observation when log-transforming the data. A total of 500 simulations were run for each sample size $n \in (500, 2500, 5000, 10000, 25000, 40000)$.

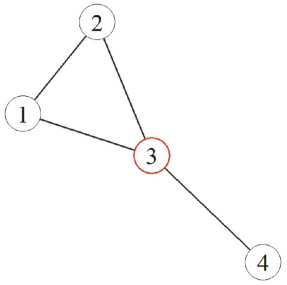


Fig. 1. Graph used for the simulations.

To address the issue of identifying differences in the network, we applied to the transformed data the *SourceSet* method (based on the normality assumption). The statistical guarantees of the tool are described in [9] and the algorithm is implemented in the `R` package `sourceSet` [10]. To improve the asymptotic approximation of the test statistics we applied the correction in [11]. Given that $\theta_3^{(1)} \neq \theta_3^{(2)}$, and following the rules in [9] we would expect that the procedure finds the node 3 as the source of the difference in the two conditions. In fact, this

algorithm identifies the smallest set containing the difference, either a clique or a separator, where a clique is a maximal subset of nodes all connected among each other, while a separator is the intersection of two cliques; we refer the reader to the book in [12] for a detailed explanation of these terms. The identification of node 3 is possible because it is a separator of the two cliques: $\{1, 2, 3\}$ and $\{3, 4\}$. Finally, FWER was controlled at 0.05 level.

3 Results

The results of the simulations are shown in Fig. 2.

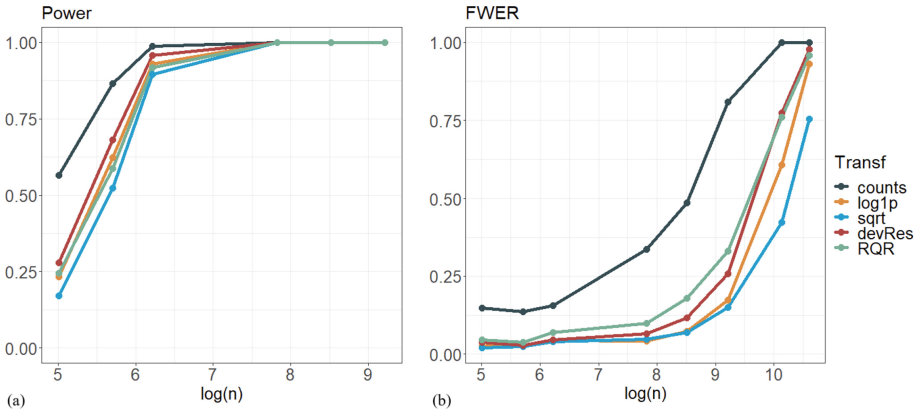


Fig. 2. (a) Power: fraction of times the node 3 is identified as different in the two conditions. (b) FWER: fraction of times at least one of the nodes 1, 2, and 4 has been falsely discovered. The black line is for reference and refers to the *raw*-counts.

Figure 2(a) shows the power, namely the fraction of times the node 3 was identified as different in the two conditions, when the procedure is applied to all the transformed data. We can observe that as the sample size n increases, the power also increases as expected. However, Fig. 2(b) shows that as n increases, also the number of false rejections increases. This behavior is common to all the transformations, with a small advantage of the square-root transformation. Nevertheless, this transformation also show an increasing trend.

Further simulations showed that, on the contrary, when the difference affects a node which is not a separator, the method has comparable results in terms of power and the FWER is controlled at the nominal level. Moreover, when considering bigger graphs with a larger number of cliques, the false rejections considering these cliques only are controlled at the nominal level. On the other hand, the same results observed in the previous paragraph are obtained for the cliques affected by the difference.

4 Discussion

When dealing with count data, assumptions about the data structures are of great importance, especially if the analysis aims to study the conditional relationships between variables. In this setting, we have highlighted the fact that the data transformation does not seem to meet the goal of the analysis. In particular, when the target object of the analysis is the detection of a difference involving a separator, as shown in Fig. 2, the *SourceSet* approach has a power increasing with the sample size, but at the same time, the number of false rejections is increasing as well. This is not observed in situations where the node affected by a change in parameter belongs to a clique only. In the same way, the number of false rejections is also controlled for all those nodes in cliques not connected to the source of the difference.

This behavior might be the consequence of the conditional dependence structure. In fact, the *SourceSet* method highly relies on the graphical structure to identify the subsets of nodes affected by a change in the two groups. Hence, the fact that data transformation might act on the structure of the graph or on the relationship between the parameters of the model, could act as a propagator of the differences in node 3, such that, in the end, all the nodes connected to it are affected. A method that is specifically designed for data coming from a normal distribution may therefore result in an unsuitable approach when applied to data of a different nature. If, indeed, the nature of the difference between the two conditions lies in a parameter θ, relative to a single node, the transformation may not necessarily maintain the relationship with the corresponding parameter of the normal distribution, in this case the mean. Additionally, in the case of count data, it is well-known that there is a dependency between the mean and variance, and undoubtedly this may affect the transformed data. Therefore, the main problem is not the ability of the test to detect the differences, but rather not knowing exactly how the difference in a certain parameter of the initial distribution is then propagated to the other parameters in the distribution of the transformed data.

5 Conclusion

Thanks to the expansion of new technologies in medical and biological fields, the amount of data available is increasing, as is their complexity. The classical statistical methods were mostly developed for analyzing data that can be assumed to come from a normal distribution. However, this work demonstrates the need to go beyond the normality assumption to chase the complexity of the data. Through a simulation study, we showed how common data transformations, whose purpose is to make data fit a normal distribution, may lead to ambiguous or even wrong conclusions. When dealing with RNA-seq data, it is important to select methods that are specifically developed for handling count data rather than methods designed to handle normally distributed data. When the focus of the analysis involves the graphical structure, which is highly reliant

on conditional dependencies among variables, the use of data transformation may alter the relationships and thus the final conclusions. We showed how the use of the transformations should be carefully considered within the context of the application.

These results highlight the need for *ad hoc* methods for handling count data in a two-sample problem. There is a distinct lack, in the literature, of alternative methods suitable for handling differential network analysis, representing an important area for future research. In particular, the KLIEP approach suggested in [13] offers promising development opportunities.

This paper concludes that future statistical research in the transcriptomic field should be focused on the development of tools designed specifically for tackling the complexity of this kind of data, while the hypothesis of transforming the data in order to use methods meant for other kinds of data should be carefully considered.

References

1. Irizarry, R.A., et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**(2), 249–264 (2003)
2. Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**(1), 57–63 (2009)
3. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., Teichmann, S.A.: The technology and biology of single-cell RNA sequencing. Mol. Cell **58**(4), 610–620 (2015)
4. Van De Wiel, M.A., Leday, G.G., Pardo, L., Rue, H., Van Der Vaart, A.W., Van Wieringen, W.N.: Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics **14**(1), 113–128 (2013)
5. Shojaie, A.: Differential network analysis: a statistical perspective. Wiley Interdisciplinary Rev. Comput. Stat. **13**(2), e1508 (2021)
6. Ahlmann-Eltze, C., Huber, W.: Comparison of transformations for single-cell RNA-seq data. Nat. Methods **20**(5), 665–672 (2023)
7. Inouye, D.I., Yang, E., Allen, G.I., Ravikumar, P.: A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdisciplinary Rev. Comput. Stat. **9**(3), e1398 (2017)
8. Zhang, R., Ren, Z., Celedón, J.C., Chen, W.: Inference of large modified Poisson-type graphical models: application to RNA-seq data in childhood atopic asthma studies. Ann. Appl. Stat. **15**(2), 831–855 (2021)
9. Djordjilović, V., Chiogna, M.: Searching for a source of difference in graphical models. J. Multivar. Anal. **190**, 1–13 (2022)
10. Salviato, E., Djordjilović, V., Chiogna, M., Romualdi, C.: SourceSet: a graphical model approach to identify primary genes in perturbed biological pathways. PLoS Comput. Biol. **15**(10), e1007357 (2019)
11. Banzato, E., Chiogna, M., Djordjilović, V., Risso, D.: A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models. Stat. Probab. Lett. **193**, 109732 (2023)
12. Lauritzen, S.L.: Graphical models. Clarendon Press (1996)
13. Liu, S., Quinn, J.A., Gutmann, M.U., Suzuki, T., Sugiyama, M.: Direct learning of sparse changes in Markov networks by density ratio estimation. Neural Comput. **26**(6), 1169–1197 (2014)

Object-Oriented Analysis of Spatial Complex Data in Stream-Network Domains

Chiara Barbi, Alessandra Menafoglio(✉), and Piercesare Secchi

MOX, Department of Mathematics, Politecnico di Milano, Milan, MI, Italy
chiara.barbi@mail.polimi.it,
{alessandra.menafoglio,piercesare.secchi}@polimi.it

Abstract. We tackle the challenge of spatial prediction for Hilbert data embedded within river networks, where the network's reticular layout mandates geostatistical approaches employing Stream Distance to capture spatial connectivity stemming from branching. Embedded within Object Oriented Spatial Statistics (O2S2), we present the innovative methodology introduced in [1]; here, the authors conceptualize data as points within a functional embedding space, and devise functional moving average models grounded in the Stream Distance. This methodology allows one to take into account both data and spatial domain geometry, enabling the establishment of a coherent covariance structure and the definition of associated estimators. By analysing summer water temperature profiles along Idaho's Middle Fork River, our approach showcases its effectiveness in modelling the covariance structure and enhancing forecasting capabilities within the O2S2 paradigm. This investigation offers valuable insights into tailored modeling techniques adept at navigating the complexities of spatially structured data in natural landscapes.

Keywords: Stream Distance · Geostatistics · moving-average models

1 Introduction

In recent years, the statistical community has faced the formidable challenge of analyzing and extracting meaningful insights from increasingly complex and diverse datasets. Traditional statistical methods designed for scalar data often fall short in contexts where data exhibit high complexity, such as curves, surfaces, or images. Consequently, Functional Data Analysis (FDA, [10]) and Object Oriented Data Analysis (OODA, [4]) have garnered significant attention in the literature, as researchers endeavor to adapt classical statistical techniques to accommodate this evolving landscape.

However, when data are georeferenced, their complexity is compounded by the necessity to account for spatial dependence induced by proximity. This has prompted a surge in literature aimed at developing spatial statistical methods

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 127–132, 2025.
https://doi.org/10.1007/978-3-031-64431-3_22

applicable to various types of data objects, including functional data, distributions, and data associated with Riemannian manifolds. This pursuit has culminated in the emergence of the so-called Object-Oriented Spatial Statistics (O2S2, [7]), a conceptual framework rooted in geometric principles for analyzing spatially complex data.

However, current methods in O2S2 primarily focus on Euclidean spatial domains or mildly non-Euclidean regions that can be locally represented in a Euclidean manner [5,6]. Yet, many areas of geosciences study random processes that naturally unfold over non-Euclidean settings, where proximity between data points is expressed through geodesic paths induced by the underlying physics. For instance, in the study of aquatic variables within a stream network, proximity among sites is better captured by their water distance rather than the Euclidean shortest path between them, which fails to account for network topology and connectivity.

In this communication, we describe the novel methodology presented in [1], where a functional moving average model using Stream Distance is established by considering both the data geometry and the topology of the spatial domain. Moreover, we illustrate a consistently developed definition of the covariance structure and examine the associated estimators. Finally, through a case study regarding summer water temperature profiles in the Middle Fork River, Idaho, we will discuss the ability of the method to effectively characterize the covariance structure and provide forecasts at unsampled locations.

2 Moving Average Models for Stream Network Data

In this study, we depict stream networks as dendritic structures composed of a finite number of stream segments. Each segment comprises point locations along a straight line, each associated with a unique water flow direction. Given the dendritic nature of the network, there is always a singular most-downstream point, henceforth termed *the outlet*. Consequently, the *upstream distance* for any point in the network represents the path length connecting it to the outlet. In the following, we denote by D the stream network domain, and by s as a generic spatial location (i.e., a point) in D. For two locations to be considered *flow-connected*, water must flow between them; otherwise, they are termed *flow-unconnected*. The stream distance is then defined as the shortest distance between two locations, with the constraint that all displacements occur along the network, that is

$$d(s_i, s_j) = \begin{cases} |s_i - s_j| & \text{if } s_i \text{ and } s_j \text{ are flow-connected,} \\ (s_i - u) + (s_j - u) & \text{otherwise.} \end{cases} \tag{1}$$

In (1), u is the nearest junction downstream which is common to both flow-unconnected locations.

In geostatistics, dealing with non-Euclidean spatial domains presents challenges due to the breakdown of typical parametric families like spherical or

Matérn, as they are no longer guaranteed to be valid when applied under a non-Euclidean metric [3]. However, in the context of stream networks, custom parametric families can be derived, tailored to the domain's topology. Extensively studied examples include the models proposed in [2,9,12], among others. These models, rooted in Yaglom's moving average construction [13], accurately capture dependence induced by water distance, offering valid covariance models and estimation procedures suitable for spatial data, when their domain can be represented as a binary tree.

In [1], we build upon the scalar framework introduced in [11] to extend the stream network methodology to functional observations. More precisely, call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and H a separable Hilbert space, with operations $(+, \cdot)$ and inner product $\langle \cdot, \cdot \rangle$, inducing the norm $\| \cdot \|$. We assume that each element of H is a function $\mathcal{X} : \tau \to \mathbb{R}$, τ being a compact subset of $\mathbb{R}$, and we call $\{\mathcal{X}_s, s \in D\}$ a functional random field valued in H, which we hereafter assume to be stationary. Note that each element of the field can be represented via a truncated basis expansion as $\mathcal{X}_s = m + \sum_{k=1}^{N} \xi_k(s) e_k$, where m denotes the (spatially-constant) mean, $\{e_k, k \geq 1\}$ is an orthonormal basis of H and the random coefficients $\xi_k(s) = \langle \mathcal{X}_s - m, e_k \rangle$ are the projections of the functional random variable $\mathcal{X}_s$ on the orthonormal basis. Notice that the stationarity and covariance structure of the functional process are determined by these coefficients. By defining each scalar field through a moving average (MA) model of the form [13]

$$\xi_k(s) = \int_{-\infty}^{+\infty} g^{(k)}(x - s|\boldsymbol{\theta}) dW_k(x), \tag{2}$$

with $g^{(k)}(\cdot|\boldsymbol{\theta})$ a (square-integrable) MA function with parameters $\boldsymbol{\theta}$, and W_k's a Wiener processes, [1] shows that the trace-covariogram of the process $\{\mathcal{X}_s, s \in D\}$ is

$$C(s_i, s_j) = \sum_{k=1}^{N} \mathbb{E}\Big[\xi_k(s_i)\xi_k(s_j)\Big] = \sum_{k=1}^{N} C_t^{(k)}(h|\boldsymbol{\theta}),$$

where $C_t^{(k)}(h|\boldsymbol{\theta})$ is the covariance function of the k-th scalar random field $\{\xi_k(s), s \in D\}$. In practice, the scalar covariance functions $C_t^{(k)}(h|\boldsymbol{\theta})$ can be modeled as one of the valid models introduced by Verhoef (2010), and so does the trace-covariogram $C(s_i, s_j)$.

3 Tail-up and Tail-down Models

Similar to the approach presented in [11], the field $\{\mathcal{X}_s, s \in D\}$ can be represented using either a tail-up or a tail-down MA model.

Tail-up models (TU) feature moving average functions $g^{(k)}(\cdot|\boldsymbol{\theta})$ that are non-null only in the upstream direction, making them suitable for variables like pollutants or waterborne chemicals, which are predominantly influenced by downstream flow. Note that the manner in which the MA function $g^{(k)}(\cdot|\boldsymbol{\theta})$ is split at network nodes while moving upstream is crucial for preserving the stationarity

of the spatial process $\{\mathcal{X}_s, s \in D\}$. Segment weights are employed to distribute the function proportionally among upstream segments when the MA function reaches a confluence in the network. In this case, covariance between elements $\mathcal{X}_{s_i}$ and $\mathcal{X}_{s_j}$ depends on flow connectivity, on the stream distance h between them and on the system of weights. In contrast, tail-down (TD) models have non-zero moving average functions only downstream of a location, allowing autocorrelation for both flow-connected and flow-unconnected locations. Unlike TU models, TD models do not require a weighting procedure at junctions, simplifying their covariance structure. Tail-down models are suitable for variables like fish or aquatic insects that can move in both directions, offering diverse covariance structures for connected and unconnected site pairs. Examples of TU and TD covariance structures are discussed in [1].

4 Model Estimate and Prediction

The estimation of covariance models is centered around the trace-semivariogram of the field. The trace-semivariogram, defined under global second-order stationarity as $\gamma(s_1, s_2) = \frac{1}{2}\mathbb{E}[\|\mathcal{X}_{s_1} - \mathcal{X}_{s_2}\|^2]$, plays a key role in characterizing the spatial dependence of the field, as it is closely related to the trace-covariogram, as per the well-known relation $\gamma(s_1, s_2) = C(s_1, s_1) - C(s_1, s_2)$.

To estimate the trace-semivariogram, the two-step approach commonly used in scalar geostatistics is employed in this setting as well. Firstly, an empirical estimator is determined, followed by fitting a suitable model to the estimated semivariogram. The models proposed in [1] are functional extension of those introduced in [12]; they are defined piecewise based on the connectedness of the pair being considered. More precisely, two empirical trace-semivariograms can be defined: *(i)* the Flow-Unconnected Stream-Distance (FUSD) and *(ii)* the Flow-Connected Stream-Distance (FCSD) trace-semivariograms. The former is computed solely from flow-unconnected site-pairs, while the latter considers flow-connected site-pairs. On this basis, [1] proposes a parameter estimation procedure relying on the empirical FCSD trace-semivariogram, which involves (*i*) estimating the empirical FCSD trace-semivariogram, (*ii*) determining the model assumption based on the FUSD semivariogram, (*iii*) fitting a valid model to the FCSD semivariogram, and (*iv*) obtaining the associated trace-covariogram. For the model identification at step (*ii*), note that flow-unconnected site-pairs are uncorrelated in pure tail-up models, yielding pure-nugget models for the flow-unconnected portion of the variogram.

Having estimated the spatial covariance, object-oriented kriging can follow the same line as that detailed in [8], provided that the estimated (stream-distance-based) trace-semivariogram is plugged in kriging system.

5 Case Study

Following [1], we illustrate the analysis of Middle Fork River water temperatures using the proposed methodology. Data were collected from July 15 to August

31, 2005, at 157 river locations, and preprocessed through spline smoothing with penalization fixed through generalized cross-validation. Elevation and upstream distance where included in the drift term to model the non-stationary behavior of the data, as suggested by the empirical trace-semivariograms. An exponential tail-up model was used to capture the spatial variability, with parameters estimated via generalized least squares. A leave-one-out cross-validation procedure confirmed that kriging prediction based on stream distance proved highly satisfactory, validating the ability of the approach to capture spatial patterns along the river network, which is crucial for environmental monitoring and management.

6 Conclusion and Discussion

In this communication, we illustrated the novel approach proposed in [1] to model functional random fields within the context of stream networks. The methodology builds upon the principles of O2S2, extending them to account for the unique characteristics of stream networks, and particularly for the possible flow-connected and flow-unconnected nature of its locations. The empirical estimation of FUSD and FCSD trace-semivariograms plays a crucial role in this setting, allowing us to capture the complex spatial dependence patterns in dendtritic spatial domain. Real data tests validate the methodology's predictive capability, yet further research is needed for tail-up models to refine semivariogram estimation by incorporating river topology weights, which are currently neglecting, yielding a (moderate) bias in the estimated covariance functions. Additional investigation is required to expand the existing models to accommodate the analysis of data dispersed across non-binary trees and general networks, where appropriate covariance models have yet to be established.

Acknowledgements. Alessandra Menafoglio and Piercesare Secchi acknowledge the support by MUR, grant Dipartimento di Eccellenza 2023–2027.

References

1. Barbi, C., Menafoglio, A., Secchi, P.: An object-oriented approach to the analysis of spatial complex data over stream-network domains. Spatial Stat. **58**, 100784 (2023)
2. Cressie, N., Frey, J., Harch, B., Smith, M.: Spatial prediction on a river network. J. Agric. Biol. Environ. Stat. **11**, 127–150 (2006)
3. Curriero, F.: On the use of non-Euclidean distance measures in geostatistics. Math. Geol. **38**, 907–926 (2006)
4. Marron, J.S., Alonso, A.M.: Overview of object oriented data analysis. Biom. J. **56**(5), 732–753 (2014)
5. Menafoglio, A., Gaetani, G., Secchi, P.: Random domain decompositions for object-oriented kriging over complex domains. Stoch. Env. Res. Risk Assess. **32**, 3421–3437 (2018)

6. Menafoglio, A., Pigoli, D., Secchi, P.: Kriging Riemannian data via random domain decompositions. J. Comput. Graph. Stat. **30**(3), 709–727 (2021)
7. Menafoglio, A., Secchi, P.: Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. Eur. J. Oper. Res. **258**(2), 401–410 (2017)
8. Menafoglio, A., Secchi, P., Dalla Rosa, M.: A universal kriging predictor for spatially dependent functional data of a Hilbert space. Electron. J. Stat. **7**, 2209–2240 (2013)
9. Peterson, E., Theobald, D., Ver Hoef, J.: Geostatistical modelling on stream networks: Developing valid covariance matrices based on hydrologic distance and stream flow. Freshw. Biol. **52**, 267–279 (2007)
10. Ramsay, J.O., Silverman, B.W.: Functional data analysis. Springer Series in Statistics. New York, NY: Springer (2005)
11. Ver Hoef, J., Peterson, E.: A moving average approach for spatial statistical models of stream networks. J. Am. Stat. Assoc. **105**, 6–18 (2010)
12. Ver Hoef, J., Peterson, E., Theobald, D.: Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat. **13**, 449–464 (2006)
13. Yaglom, A.M.: Correlation theory of stationary and related random functions, vol. 1 of Springer Series in Statistics,. New York, NY: Springer-Verlag (1987)

Latent Class Multi-state Quantile Regression

Rosario Barone(✉) and Alessio Farcomeni

University of Rome "Tor Vergata", Rome 00133, Italy
rosario.barone@uniroma2.it

Abstract. We propose a multi-state quantile regression model that admits a cure-fraction for each possible transition, so that individuals may not experience that event. A discrete latent variable allows us to take into account unobserved heterogeneity. The model is estimated in a Bayesian framework, without specification of the number of latent classes. We are motivated by an original application to spells of imprisonment in the USA.

Keywords: Asymmetric Laplace Distribution · Cure rate models · Dirichlet Process prior

1 Introduction and Set-Up

Consider a stochastic process X_t for $t \geq 0$ that evolves in a discrete state-space $\{1, \ldots, s\}$ where $s \geq 2$ and continuous time. Assume that state m is not an absorbing state. Let $T_m = \inf\{t \geq 0 : X_t \neq m, X_0 = m\}$ be the random variable representing the time of the first jump from state m. Additionally, let Y_j denote the sequence of states visited by the continuous-time Markov chain (regardless of dwelling times), with Y_0 as the initial state. Thanks to the Markov homogeneity assumptions, the stochastic process is entirely determined by

$$F_{mj}(t) = \Pr(T_m \leq t, Y_1 = j | Y_0 = m) = \Pr(T_{mj} \leq t), \quad t > 0 \tag{1}$$

where T_{mj} is the time taken to transition from state m to state j, $m = 1, \ldots, s$, and $j \neq m$. The variable T^*_{mj} is defined as $T^*_{mj} = T_{mj}$ if the transition is to state j and $T^*_{mj} = \infty$ if the transition is to some other state. When the follow-up is closed in state m, there is an independent censoring time C_m, where $C_m = \infty$ if a transition occurs. By defining the $p \times 1$ vector of predictors $X = (X_1, X_2, \ldots, X_p)'$, Farcomeni and Geraci (2020) model the conditional quantile function

$$Q_{mj}(\tau | X) = \inf\{t : \Pr(T^*_{mj} \leq t, Y_1 = j | X, Y_0 = m) \geq \tau\}, \tag{2}$$

$\tau \in (0, 1)$, which is the left inverse of the conditional CDF $\Pr(T^*_{mj} \leq t, Y_1 = j | X, Y_0 = m)$, as

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 133–138, 2025.
https://doi.org/10.1007/978-3-031-64431-3_23

$$Q_{mj}(\tau|X) = g_{mj}\{\alpha(\tau) + X'\beta_{mj}(\tau)\}, \tag{3}$$

where $g_{mj}(\cdot)$ represents a known monotone link function, $\alpha(\tau)$ as a quantile-specific intercept, and $\beta_{mj}(\tau) = (\beta_{mj,1}(\tau), \beta_{mj,2}(\tau), \ldots, \beta_{mj,p}(\tau))'$ a vector of quantile-specific regression coefficients. We let $\lim_{t\to\infty}(1 - F_{mj}(t)) \geq 0$, that is, for a fraction of individuals to possibly never experience the transition from m to j. We thus provide a quantile-regression counterpart of Conlon et al. (2014), that model instantaneous hazards in the presence of a cure fraction. We then further extend the model to allow for unobserved heterogeneity, that is captured through a discrete latent variable. Our Bayesian inference framework allows us not specify the number of latent classes, and provides a very efficient inferential framework that scales well also for large data sets. We conclude with an original application to spells of imprisonment in the United States of America.

2 Cured Fraction in MSM Quantile Regression

We introduce the binary latent variable R_{mj}, with $R_{mj} = 1$ if for all t

$$\Pr(T_{mj} \leq t|R_{mj} = 1) = 0. \tag{4}$$

Let also U denote another latent variable, with $\Pr(U = k) = \pi_k$, which will be used to cluster subjects. We model

$$\begin{cases} \Pr(R_{mj} = 1|Z, U) = \eta_{mj}\left(\gamma_{0U} + Z'\gamma_{mj}\right) \\ Q_{mj}(\tau|X, U, R_{mj} = 0) = g_{mj}\{\beta_{0U}(\tau) + X'\beta_{mj}(\tau)\}. \end{cases}$$

where Z represents a matrix of covariates, $\eta_{mj}(\cdot)$ is a link function for binary regression models, and γ_{mj} is a parameter vector explaining the effect of Z on the probability of being cured. In words, intercepts are not homogeneous. Subjects belonging to the same latent class share a common value for the intercept, which can be interpreted as common propensities to transitioning from state m to state j or to stay in state m indefinitely.

We obtain a likelihood framework by specifying Asymmetric Laplace Densities (ALD) for t (Yu and Moyeed, 2001). For $i = 1, \ldots, N$, let $\mathbf{t}_i^{(mj)} = (t_{i1}^{(mj)}, \ldots, t_{in_i}^{(mj)})$ be the set of observed transition times between the states m and j, with n_i representing the total number of transitions from m to j observed for the individual i, and let $T_{mj} = \left(\mathbf{t}_1^{(mj)}, \ldots, \mathbf{t}_N^{(mj)}\right)$ be the set containing all the observed transition times from m to j across the N observed individuals. Let the vector of observed covariates be $\left\{X_{i1}^{(mj)}, \ldots, X_{in_i}^{(mj)}\right\}$ and $\left\{Z_{i1}^{(mj)}, \ldots, Z_{in_i}^{(mj)}\right\}$, with $X_{il}^{(mj)}$ of length d and $Z_{il}^{(mj)}$ of length p for $l = 1, \ldots, n_i$. We consider $c_{mj}^{(il)} = 1$ if the l-th observation of the i-th individual is censored and $c_{mj}^{(il)} = 0$ if

the observation is uncensored. The observed data likelihood when $k = 1$ can be written as

$$\mathcal{L}(\beta_{mj}, \sigma_{mj}, \gamma_{mj}) = \prod_{i:n_i=1} \left\{ \left[\left(1 - \eta_{mj}\left(Z_i^{(mj)'}\gamma_{mj}\right)\right) f\left(t_{i1}^{(mj)} | \beta_{mj}, \sigma_{mj}, \tau, X_{i1}^{(mj)}\right)\right]^{\left(1-c_{mj}^{(i1)}\right)} \right. \tag{5}$$

$$\left. \left[\eta_{mj}\left(Z_i^{(mj)'}\gamma_{mj}\right) + \left(1 - \eta_{mj}\left(Z_i^{(mj)'}\gamma_{mj}\right)\right) S\left(t_{i1}^{(mj)} | \beta_{mj}, \sigma_{mj}, \tau, X_{i1}^{(mj)}\right)\right]^{c_{mj}^{(i1)}} \right\}$$

$$\prod_{i:n_i>1} \left(1 - \eta_{mj}\left(Z_i^{(mj)'}\gamma_{mj}\right)\right) \prod_{l=1}^{n_i} \left\{ f\left(t_{il}^{(mj)} | \beta_{mj}, \sigma_{mj}, \tau, X_{il}^{(mj)}\right)^{\left(1-c_{mj}^{(il)}\right)} \right.$$

$$\left. S\left(t_{il}^{(mj)} | \beta_{mj}, \sigma_{mj}, \tau, X_{il}^{(mj)}\right)^{c_{mj}^{(il)}} \right\}$$

where $f(\cdot)$ denotes the ALD density and $S(\cdot)$ denotes the ALD survival function, namely $S(\cdot) = 1 - F(\cdot)$.

We rely on an augmentation framework when $k > 1$. By defining $\theta_{mj} = (\beta_{mj}, \sigma_{mj}, \gamma_{mj})$ as the vector containing the model parameters constant among the latent classes and $\alpha_{U_{mj}} = (\beta_{U_{0mj}}, \gamma_{U_{0mj}}) \in A$ as the vector of the parameters that vary among latent classes, we define a mixture model with a kernel p, defined as the joint density of transitions times from the state m to j for the N observed individuals. Formally, let G be a probability distribution defined on the parameter space A. Let $\mathrm{DP}(MG_0)$ denote a Dirichlet Process with base measure G_0 and scaling parameter M. With a DP prior on the mixing measure G, we get a Dirichlet Process Mixture (DPM), which may be represented as

$$p_{\alpha_{mj}}(\mathbf{t}_i^{(mj)} | \theta_{mj}, \tau, X, Z, c) = \sum_{k=0}^{\infty} w_k p_{\alpha_{k_{mj}}}(\mathbf{t}_i^{(mj)} | U = k, \theta_{mj}, \tau, X, Z, c), \tag{6}$$

where w_k are mixture weights such that $\sum_{k=0}^{\infty} w_k = 1$. The posterior distribution $\Pi(G|\theta_{mj}, \tau, T_{mj}, X, Z, c)$ is a mixture of DP models, mixing with respect to latent variables $\alpha_{i_{mj}}$ specific to each observed individual $\mathbf{t}_i$ for $i = 1, \ldots, N$:

$$G|\theta_{mj}, \tau, T, X, Z, c \sim \int DP(MG_0 + \sum_{i=1}^{n} \delta_{\alpha_{i_{mj}}}) d\Pi(\alpha_{mj} | \theta_{mj}, \tau, T_{mj}, X, Z, c),$$

with $\alpha_{mj} = (\alpha_{1_{mj}}, \ldots, \alpha_{N_{mj}}) = ((\beta_{01}, \gamma_{01}), \ldots, (\beta_{0N}, \gamma_{0N}))$. DPMs provide a way to model the uncertainty about the number of components in the mixture and their parameters, automatically determining the appropriate number of clusters given the data. This feature ensures the adequate flexibility of the model, allowing it to adapt to the characteristics of the observed individuals.

The infinite dimensional mixture model is managed by following Walker (2007) and Kalli et al. (2011), that proposed a slice sampler for DPM models, introducing uniform auxiliary variables which make the mixture model conditionally finite. The overall sampling scheme employs an MCMC algorithm, wherein we classify individual observations and update the posterior distribution of cluster parameters α. Finally, we update the vector $\theta = (\beta, \sigma, \gamma)$. At iteration $h + 1$ of the MCMC sampler, for each state m and j with $m \neq j$:

– define the number of components K^* via slice sampling;
– update individual labels: for $k = 1, \ldots, K^*$
  draw $\psi_i^{(h+1)} \sim \pi(\psi_i = k | \alpha_k^{(h)}, \theta^{(h)}, \tau, \mathbf{t}_i, X, Z, c_i)$
– update the label dependent vector of parameters α: for $k = 1, \ldots, K^*$
  draw $\alpha_k^{(h+1)} \sim \pi(\alpha_k | \beta^{(h)}, \sigma^{(h)}, \gamma^{(h)}, \tau, T, X, Z, c, k^{(h+1)})$
– update the label independent vector of parameters:
  draw $(\beta, \sigma, \gamma)^{(h+1)} \sim \pi(\beta, \sigma, \gamma | \alpha^{(h+1)}, \tau, T, X, Z, c)$.

3 Analysis of Jail Data

We implemented the proposed model on a sample of 8143 individuals who were incarcerated at least once across the USA, for a total of 19932 observations. We have $s = 2$, with the first arrest serving as the starting time for each observed individual. Additionally, a large number of covariates, specified in Table 1, were considered, and these were utilized for both the multi-state quantile regression and the cure rate model. Our model accounts for two potential transitions: from the prisoner state to the freedom state, and vice versa, each modeled independently. Given the inconsistency of DPM models in estimating mixture components (Miller and Harrison, 2013), we adopt the approach of Frühwirth-Schnatter and Malsiner-Walli (2019), setting the concentration parameter M = 1/20 to achieve a reliable estimator for latent class numbers. We defined the base measure as a bivariate Normal distribution centered on zero and fixed a quantile level $\tau = 0.5$. As shown in Table 2, at convergence we obtain $\hat{k} = 1$ for the transition from "freedom" to "prisoner" state, while $\hat{k} = 5$ in the transition from "prisoner" to "freedom" state. The majority of observations have fast transitions to the freedom state, as individuals are often arrested for non-serious offenses and released shortly after, typically upon payment of bail. For the multi-state quantile regression, higher values of the linear combination correspond to a lower tendency to exit the state, whereas - for the cured rate model - the higher the value of the linear combination, the greater the probability of remaining indefinitely in the occupied state. Estimates of coefficients related to the type of

Table 1. Covariates used in the analysis and their numerical code for Table.

1 Age	9 Conspiracy to Commit Murder	17 Rape
2 Sex	10 Distribution Cocaine or Crack	18 Weapons Offense
3 Aggravated Assault	11 Distribution Drug Unspecified	19 Black
4 Armed Robbery	12 Forgery Fraud	20 White
5 Arson	13 Fraudulent Drug Offense	21 Hispanic
6 Attempted Murder	14 Grand Theft over 500	22 Asian/Pacific Americans
7 Child Abuse	15 Kidnapping	23 Native Americans
8 Child Molestation	16 Murder	

Table 2. Summary statistics of the posterior distributions of the model parameters for quantile level $\tau = 0.5$ for both transitions.

Freedom to Prisoner												
Multi-state quantile regression	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
$E(\cdot\|T,X,Z)$	−0.005	0.057	0.076	−0.502	−0.119	−0.210	−0.233	0.002	0.458	0.442	−0.092	−0.279
$SD(\cdot\|T,X,Z)$	0.001	0.003	0.002	0.006	0.008	0.004	0.012	0.003	0.010	0.009	0.005	0.005
$q_{0.025}(\cdot\|T,X,Z)$	−0.006	0.051	0.071	−0.513	−0.132	−0.219	−0.256	−0.003	0.438	0.421	−0.103	−0.287
$q_{0.975}(\cdot\|T,X,Z)$	−0.003	0.064	0.081	−0.490	−0.101	−0.201	−0.206	0.007	0.477	0.458	−0.081	−0.270
	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}	β_{21}	β_{22}	β_{23}	
$E(\cdot\|T,X,Z)$	−0.063	0.041	0.354	−0.285	−0.002	−0.274	−0.182	−0.149	−0.297	0.620	0.383	
$SD(\cdot\|T,X,Z)$	0.004	0.008	0.005	0.008	0.003	0.007	0.010	0.005	0.009	0.008	0.010	
$q_{0.025}(\cdot\|T,X,Z)$	−0.072	0.024	0.343	−0.300	−0.007	−0.286	−0.199	−0.159	−0.314	0.604	0.362	
$q_{0.975}(\cdot\|T,X,Z)$	−0.054	0.056	0.364	−0.266	0.004	−0.261	−0.160	−0.137	−0.276	0.634	0.402	
Cure rate model	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}
$E(\cdot\|T,X,Z)$	0.013	0.326	−0.199	0.315	0.562	0.311	−0.022	0.290	0.243	0.013	−0.222	−0.176
$SD(\cdot\|T,X,Z)$	0.001	0.009	0.006	0.005	0.007	0.009	0.004	0.012	0.005	0.008	0.005	0.006
$q_{0.025}(\cdot\|T,X,Z)$	0.011	0.305	−0.209	0.306	0.549	0.292	−0.030	0.262	0.234	−0.003	−0.231	−0.188
$q_{0.975}(\cdot\|T,X,Z)$	0.015	0.342	−0.185	0.323	0.575	0.326	−0.015	0.311	0.253	0.028	−0.209	−0.164
	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{17}	γ_{18}	γ_{19}	γ_{20}	γ_{21}	γ_{22}	γ_{23}	
$E(\cdot\|T,X,Z)$	−0.144	0.255	−0.106	−0.146	−0.472	−0.173	−0.271	−0.219	−0.196	0.100	0.540	
$SD(\cdot\|T,X,Z)$	0.002	0.014	0.003	0.003	0.010	0.007	0.008	0.008	0.004	0.006	0.009	
$q_{0.025}(\cdot\|T,X,Z)$	−0.148	0.229	−0.112	−0.152	−0.489	−0.185	−0.288	−0.234	−0.203	0.087	0.523	
$q_{0.975}(\cdot\|T,X,Z)$	−0.140	0.285	−0.099	−0.140	−0.451	−0.158	−0.255	−0.204	−0.185	0.111	0.557	
		$k=1$										
Intercepts	β_0	γ_0										
$E(\cdot\|T,X,Z)$	0.778	0.820										
$SD(\cdot\|T,X,Z)$	0.017	0.021										
$q_{0.025}(\cdot\|T,X,Z)$	0.744	0.772										
$q_{0.975}(\cdot\|T,X,Z)$	0.815	0.855										
Prisoner to Freedom												
Multi−state quantile regression	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
$E(\cdot\|T,X,Z)$	−0.001	0.032	−0.167	0.256	−0.057	−0.204	0.273	0.286	−0.571	−0.108	−0.139	−0.062
$SD(\cdot\|T,X,Z)$	0.000	0.001	0.003	0.001	0.005	0.002	0.002	0.002	0.005	0.002	0.002	0.001
$q_{0.025}(\cdot\|T,X,Z)$	−0.001	0.030	−0.175	0.254	−0.068	−0.207	0.271	0.280	−0.582	−0.111	−0.143	−0.064
$q_{0.975}(\cdot\|T,X,Z)$	−0.001	0.035	−0.162	0.258	−0.047	−0.200	0.278	0.291	−0.558	−0.104	−0.136	−0.059
	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}	β_{21}	β_{22}	β_{23}	
$E(\cdot\|T,X,Z)$	−0.019	0.148	0.774	0.366	0.161	−0.122	0.408	0.328	−0.046	−0.824	0.290	
$SD(\cdot\|T,X,Z)$	0.008	0.007	0.002	0.002	0.001	0.001	0.005	0.003	0.004	0.004	0.003	
$q_{0.025}(\cdot\|T,X,Z)$	−0.035	0.133	0.770	0.362	0.160	−0.125	0.399	0.323	−0.055	−0.834	0.284	
$q_{0.975}(\cdot\|T,X,Z)$	−0.002	0.163	0.780	0.370	0.163	−0.119	0.418	0.333	−0.037	−0.816	0.298	
Cure rate model	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}
$E(\cdot\|T,X,Z)$	0.138	−0.327	−0.190	−0.181	−0.011	−0.001	0.407	0.485	0.658	0.371	−0.461	−0.084
$SD(\cdot\|T,X,Z)$	0.001	0.002	0.005	0.007	0.004	0.007	0.004	0.002	0.004	0.003	0.002	0.003
$q_{0.025}(\cdot\|T,X,Z)$	0.136	−0.332	−0.199	−0.194	−0.018	−0.015	0.401	0.480	0.652	0.366	−0.467	−0.092
$q_{0.975}(\cdot\|T,X,Z)$	0.139	−0.322	−0.179	−0.166	−0.003	0.011	0.415	0.490	0.667	0.377	−0.456	−0.078
	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{17}	γ_{18}	γ_{19}	γ_{20}	γ_{21}	γ_{22}	γ_{23}	
$E(\cdot\|T,X,Z)$	0.346	0.311	0.055	0.384	0.001	−0.322	−0.758	−0.235	−0.693	−0.085	0.101	
$SD(\cdot\|T,X,Z)$	0.002	0.007	0.001	0.002	0.001	0.004	0.005	0.001	0.003	0.006	0.006	
$q_{0.025}(\cdot\|T,X,Z)$	0.343	0.297	0.052	0.380	−0.002	−0.331	−0.770	−0.237	−0.699	−0.096	0.088	
$q_{0.975}(\cdot\|T,X,Z)$	0.352	0.326	0.057	0.389	0.003	−0.315	−0.750	−0.233	−0.686	−0.071	0.111	
		$k=1$		$k=2$		$k=3$		$k=4$		$k=5$		
Intercepts	β_0	γ_0	β_0	γ_0	β_0	γ_0	β_0	γ_0	β_0	γ_0		
$E(\cdot\|T,X,Z)$	−0.128	−8.897	3.326	−5.620	5.940	−6.359	1.875	0.152	9.392	−24.072		
$SD(\cdot\|T,X,Z)$	0.004	0.254	0.089	0.724	0.141	1.165	0.187	1.193	0.923	6.127		
$q_{0.025}(\cdot\|T,X,Z)$	−0.135	−9.428	3.117	−7.059	5.637	−9.181	1.556	−2.187	8.232	−29.446		
$q_{0.975}(\cdot\|T,X,Z)$	−0.119	−8.505	3.482	−4.485	6.248	−4.690	2.228	1.691	10.419	−0.922		
Mixture weights	w_1	w_2	w_3	w_4	w_5							
$E(\cdot\|T,X,Z)$	0.9891	0.0074	0.0013	0.0020	0.0002							
$SD(\cdot\|T,X,Z)$	0.0011	0.0009	0.0005	0.0006	0.0002							
$q_{0.025}(\cdot\|T,X,Z)$	0.9869	0.0000	0.0000	0.0000	0.0000							
$q_{0.975}(\cdot\|T,X,Z)$	1.0000	0.0092	0.0024	0.0031	0.0006							

charge show a higher tendency for recidivism and an higher tendency to remain in prison among individuals committing violent crimes. Regarding ethnicity, there is a lower tendency for recidivism among Asian/Pacific Americans and Native Americans. The latter also exhibit a higher tendency to remain indefinitely in the freedom state.

References

Farcomeni, A., Geraci, M.: Multistate quantile regression models. Stat. Med. **39**(1), 45–56 (2020)

Conlon, A.S.C., Taylor, J.M.G., Sargent, D.J.: Multi-state models for colon cancer recurrence and death with a cured fraction. Stat. Med. **33**(10), 1750–1766 (2014)

Yu, K., Moyeed, R.A.: Bayesian quantile regression. Stat. Probab. Lett. **54**(4), 437–447 (2001)

Rosenthal, J.S.: Optimal proposal distributions and adaptive MCMC. Handb. Markov Chain Monte Carlo, **4**(10.1201) (2011)

Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat. Simul. Comput. **36**(1), 45–54 (2007)

Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. **21**, 93–105 (2011)

Frühwirth-Schnatter, S., Malsiner-Walli, G.: From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. Adv. Data Anal. Classif. **13**, 33–64 (2019)

Miller, J.W., Harrison, M.T.: A simple example of Dirichlet process mixture inconsistency for the number of components. Adv. Neural Inf. Proce. Syst. 26 (2013)

ChatGPT and Conversational AI in Customer Satisfaction: A Preliminary Literature Review

Elena Barzizza, Nicolò Biasetton(✉), and Luigi Salmaso

Department of Management and Engineering, Universitá degli Studi di Padova, Stradella San Nicole 3, 36100 Vicenza, Italy
elena.barzizza@phd.unipd.it,
{nicolo.biasetton,luigi.salmaso}@unipd.it

Abstract. Chatbots powered by Generative Pre-trained Transformers (GPT) have emerged as a powerful tool in various domains, including customer satisfaction analysis. This work presents a preliminary literature review of the adoption of ChatGPT in the context of customer satisfaction. This preliminary work It examines existing research and literature reviews on the use of ChatGPT for customer satisfaction analysis with a focus on the methodologies and findings of sentiment analysis. Our review aims to provide insights into the effectiveness, challenges, and future directions of employing ChatGPT to enhance customer satisfaction analysis procedures. We also highlight research paths emphasized by fellow scholars.

Keywords: ChatGPT · Customer Satisfaction · Customer Experience · CX · Smart data

1 Introduction and Background

Customer satisfaction stands as a cornerstone of business success, shaping brand loyalty, customer retention, and overall profitability. In the era of digital transformation, businesses are continually seeking innovative approaches to enhance their customer service processes and better understand the sentiments and preferences of their clients. The evolution of customer service from traditional interactions to digital platforms has been marked by significant milestones. Early attempts at automating customer interactions through rule-based chatbots paved the way for more sophisticated conversational agents capable of understanding natural language and context. Advancements in Natural Language Processing (NLP) technologies have played a fundamental role in enhancing the capabilities of chatbots, enabling them to engage in more human-like conversations and provide personalized assistance. This evolution can be attributed to groundbreaking research initiatives such as the introduction of Generative Pre-trained Transformers (GPT), a seminal work by Radford et al. [2018], which

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 139–144, 2025.
https://doi.org/10.1007/978-3-031-64431-3_24

has served as a cornerstone in advancing generative AI capabilities. GPT models, built upon transformer architectures (Vaswani et al. [2017]), represent a paradigm shift in NLP, leveraging large-scale pre-training on diverse text corpora to acquire a deep understanding of language structure and semantics. Subsequent fine-tuning for specific tasks - exemplified by models like ChatGPT, introduced by OpenAI in November 2023, that has gathered huge attention for its ability to generate human-like text - has further refined their capabilities, enabling them to generate coherent and contextually relevant responses to a wide array of prompts. This progress has been fueled by a confluence of factors, including advancements in deep learning techniques, availability of massive datasets, and innovations in model architecture. Notably, Adiwardana et al. [2020] demonstrated significant strides towards human-like open-domain chatbots, underscoring the transformative potential of generative AI in enabling more natural and engaging human-computer interactions. As researchers continue to push the boundaries of generative AI, leveraging novel approaches and addressing inherent challenges, the field is poised for continued growth and innovation. The present preliminary work, which focuses on analyzing literature to provide insights into the effectiveness, challenges, and future directions of employing ChatGPT in customer satisfaction studies, is organized as follows: Sect. 2 presents the methodology of the literature review, and Sect. 3 reports the main results that are discussed and from which some conclusions are drawn in Sect. 4.

2 Methodology

In this section we describe the adopted literature review approach. The ultimate aim of this review is to understand what the state of the art is in the application of chatGPT in the customer satisfaction field. To this end, three research queries were put to the Scopus database, most recently updated in February 2024:

Search 1 - Concerning ChatGPT. The research keyword was "chatgpt". The final research query was:

"*chatgpt*"

This was applied to titles, keywords and abstracts. The research results were refined by selecting the following Subject Areas: Computer Science, Engineering, Social Science, Decision Science, Business Management and Accounting. The results were further refined by selecting Keywords: Semantics and NLP. Only documents in English were selected. The final sample comprised 137 documents.

Search 2 - Concerning ChatGPT and Customer satisfaction The research keywords were "chatgpt", "llm", "large language model", "chatbot", "generative AI", "conversational AI", "genAI", "customer satisfaction", "customer experience", "customer journey", "CX"[1] , and "literature review". The final query was:

[1] Frequently adopted in Customer Satisfaction literature, stands for Customer eXperience.

"(*chatgpt* OR *llm* OR *large language model* OR *chatbot* OR *generative ai* OR *conversational ai* OR *genai*) AND (*consumer satisfaction* OR *customer experience* OR *customer journey* OR *cx*) AND *literature review*"

This was applied to titles, keywords and abstracts.The results were refined by selecting only the following Document types: Article, Conference paper, Conference review and Review. Only documents in English were selected and the final dataset comprised 21 documents.

Search 3 - Concerning ChatGPT and Sentiment analysis The research keywords were "chatgpt","sentiment analysis", "aspect based sentiment analysis", "NLP", and "context*"[2]. The final query was:

"*chatgpt* AND (*sentiment analysis* OR *aspect base sentiment analysis* OR *NLP*) AND *context**"

This was applied to titles, keywords and abstracts. Only documents in English were selected and the final dataset comprised 87 documents.

The results were combined into a single dataset and after reading abstracts, the dataset was reduced to 31 documents. After snowballing and reading the full documents, 18 were found to be literature reviews of interest as a preliminary path to understanding the adoption of chatGPT in the context of customer satisfaction.

3 Exploratory Literature Insights

This section offers a broad overview of the literature review's main insights, comprising the current state of research on ChatGPT in the customer satisfaction field, highlighting its potential applications, challenges, and areas for improvement Table 1.

4 Discussion and Conclusions

The exploration of ChatGPT and its impact on various domains has unveiled several key themes and concerns. One prominent aspect is the emphasis on moderating factors and environmental costs associated with the deployment of AI technologies. While ChatGPT offers promising applications in fields like customer service and content generation, there's a growing awareness of the need to mitigate its environmental footprint and consider ethical implications. This represents a research direction worthy of exploration and analysis. Moreover, the versatility of ChatGPT is evident in its different applications, ranging from customer service to content generation. However, alongside these opportunities come concerns about perpetuating biases and generating harmful content. As AI systems like ChatGPT become more pervasive in our daily lives, addressing

[2] The final asterisk is used to indicate all words that begin with what precedes the asterisk: in the specific case also words as "contextual" are searched.

Table 1. Literature Review Summary on ChatGPT in Customer Satisfaction

Reference	Insights
Abdelkader [2023]	Examines gaps in ChatGPT's moderating role in customer experience in digital marketing, emphasizes the need to study moderating factors.
Alshami et al. [2023]	Propose ChatGPT usage for automating systematic reviews, highlighting current and future applications such as natural language processing, chatbots, dialogue systems, and speech recognition.
Baber et al. [2023]	Provides a systematic and bibliometric review of ChatGPT literature, emphasizes the importance of considering environmental costs of GenAI and its application across various sectors.
Barreto et al. [2023]	Describe opportunities and challenges of large language models, highlighting ChatGPT's usage and its potential impacts, including equity, misuse, and environmental costs.
Bandi [2023]	Offers an overview of requirements, models, and challenges of generative AI, focusing on expanding model taxonomy and developing standardized evaluation metrics.
Belgacem et al. [2023]	Provides a comprehensive analysis of ChatGPT architecture and potential applications, including data processing in data science.
Goar et al. [2023]	Offer a generic review of the ChatGPT model with a focus on application fields such as customer service and content generation, highlight the possibility of perpetuating biases and generating harmful content.
Javaid et al. [2023]	Investigate ChatGPT usage in Industry 4.0, highlighting the need to improve context comprehension, common sense reasoning, and bias reduction.
Kar et al. [2023]	Examine the impact of GenAI in various functions and industrial sectors.
Khennouche et al. [2024]	Offer an extensive review of FAQ answering using ChatGPT while explore ChatGPT usage in responding to customer complaints.
Lin et al. [2023]	Analyze AI-driven chatbot implementation methods over the past 20 years, emphasizing the need to develop emotionally empathetic chatbots.
Omidvar and An [2023]	Propose a semantic in-context learning approach for conversational agents, highlighting challenges in using proprietary data and maintaining model updates without fine-tuning.
Roumeliotis and Tselikas [2023]	Provide a detailed preliminary review of ChatGPT training and applications, with emphasis on adaptation needs across various fields.
Sohail et al. [2023]	Offer an overview of current ChatGPT research, including application areas, multimodal data, and challenges, with a detailed table of the latter.
Tudi et al. [2023]	Combine aspect-based sentiment analysis with ChatGPT to enhance model performance, focusing on context-based learning.
Zhang et al. [2024]	Explore aspect-based sentiment quadruple extraction, including ChatGPT contributions, providing a comprehensive and thorough review.

these ethical concerns becomes paramount to ensure responsible deployment. Such path, although recognized as very important, is not the main focus of the present work but will be explored in further research. Context comprehension and bias reduction thus emerge as critical areas for improvement in ChatGPT's functionality. While it excels in many tasks, ensuring it understands nuanced contexts and minimizes biases is essential for fostering trust and reliability in its outputs. Additionally, the management-focused discussions surrounding ChatGPT highlight the need for holistic approaches that consider not only technical capabilities but also strategic and ethical implications.

Challenges in using proprietary data and updating models without fine-tuning further underscore the complexity of integrating AI technologies like ChatGPT into existing systems. These challenges necessitate innovative solutions to balance data privacy concerns with the need for continuous model improvement and adaptation. Customer satisfaction emerges as a central concern in the application of ChatGPT. While the technology holds promise for enhancing customer experiences in various domains, including customer service and content generation, its successful implementation hinges on addressing key challenges and considerations. Efforts to improve context comprehension, mitigate biases, and ensure responsible content generation are critical to fostering positive interactions and maintaining high levels of customer satisfaction. Additionally, the need for continuous improvement and relevance in ChatGPT's responses underscores the importance of ongoing research and development efforts. By prioritizing customer satisfaction and incorporating ethical considerations into the development and deployment of ChatGPT, businesses and organizations can leverage AI technologies to build stronger relationships with their customers and drive long-term success.

Acknowledgments. Nicolò Biasetton and Luigi Salmaso thank MICS (Made in Italy – Circular and Sustainable) Extended Partnership. Funding was received from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11–10-2022, PE00000004).

References

Abdelkader, O.A.: ChatGPT's influence on customer experience in digital marketing: investigating the moderating roles. Heliyon **9**(8), e18770 (2023)

Adiwardana, D., et al.: Towards a human-like open-domain chatbot. *arXiv preprint*arXiv:2001.09977 (2020)

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A.E.E., Zayed, T.: Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems **11**(7), 351 (2023)

Baber, H., Nair, K., Gupta, R., Gurjar. K.: The beginning of ChatGPT – a systematic and bibliometric review of the literature. Information and Learning Sciences (2023)

Bandi, A.: The power of generative AI: a review of requirements, models. evaluation metrics, and challenges. Future Int. Input-Output Formats **15**(8), 260 (2023)

Barreto, F., Moharkar, L., Shirodkar, M., Sarode, V., Gonsalves, S., Johns, A.: Generative artificial intelligence: opportunities and challenges of large language models. In: Balas, V.E., Semwal, V.B., Khandare, A., eds, Intelligent Computing and Networking, vol. 699, pp. 545–553. Springer Nature Singapore, Singapore, Series Title: Lecture Notes in Networks and Systems (2023)

Belgacem, A., Bradai, A., Beghdad-Bey, K.: ChatGPT backend: a comprehensive analysis. In: 2023 International Symposium on Networks, Computers and Communications (ISNCC), pp. 1–6, Doha, Qatar, IEEE (2023)

Goar, V., Yadav, N.S., Yadav, P.S.: Conversational AI for natural language processing: an review of ChatGPT. Int. J. Recent Innovation Trends Comput. Commun. **11**(3s), 109–117 (2023)

Javaid, M., Haleem, A., Singh, R.P.: A study on ChatGPT for industry 4.0: background, potentials, challenges, and eventualities. J. Econ. Technol. **1**, 127–143 (2023)

Kar, A.K., Varsha, P.S., Rajan, S.: Unravelling the impact of generative artificial intelligence (GAI) in industrial applications: a review of scientific and grey literature. Glob. J. Flex. Syst. Manag. **24**(4), 659–689 (2023)

Khennouche, F., Elmir, Y., Himeur, Y., Djebari, N., Amira, A.: revolutionizing generative pre-traineds: insights and challenges in deploying chatgpt and generative chatbots for FAQs. Expert Syst. Appl. **246**, 123224 (2024)

Lin, C.-C., Huang, A.Y.Q., Yang, S.J.H.: A Review of AI-driven conversational chatbots implementation methodologies and challenges (1999–2022). Sustainability **15**(5), 4012 (2023)

Omidvar, A., An, A.: Empowering conversational agents using semantic in-context learning. In: Proceedings of the 18th Workshop on Innovative use of NLP for Building Educational Applications (BEA 2023), pp. 766–771,Toronto, Canada (2023). Association for Computational Linguistics

Radford, A., et al.: Improving language understanding by generative pre-training (2018)

Roumeliotis, K.I., Tselikas, N.D.: ChatGPT and Open-AI Models: a preliminary review. Future Int. **15**(6), 192 (2023)

Sohail, S.S., et al.: Decoding ChatGPT: a taxonomy of existing research, current challenges, and possible future directions. J. King Saud Univer. Comput. Inf. Sci. **35**(8), 101675 (2023)

Tudi, M.R., Na, J.C., Liu, M., Chen, H., Dai, Y., Yang, L.: Aspect-based sentiment analysis of racial issues in singapore: enhancing model performance using ChatGPT. In: Goh, D.H., Chen, S.J., Tuarob, S., eds, Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration, vol. 14457, pp. 41–55. Springer Nature Singapore, Singapore (2023). Series Title: Lecture Notes in Computer Science

Vaswani, A.: Attention is all you need. Adv. Neural Inf. Proce. Syst. 30 (2017)

Zhang, H., Cheah, Y.-N., Alyasiri, O.M., An, J.: Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey. Artif. Intell. Rev. **57**(2), 17 (2024)

Eco-Sustainable Tourism in Europe, an Analysis with Multilevel Models

Francesca Bassi[1(✉)] and Juan Antonio Marmolejo Martìn[2]

[1] University of Padova, Padova, Italy
francesca.bassi@unipd.it
[2] University of Granada, Granada, Spain
jamarmo@ugr.es

Abstract. The objective of this paper is to understand which characteristics influence the decisions, among European citizens, to adopt eco-sustainable behavior during their holidays. The data comes from the Flash Eurobarometer 499 survey, carried out in 2021 with the participation of 25,714 European citizens from the 27 Member States. The questionnaire includes ten questions about participants' intentions to implement various eco-sustainable actions. Citizens' preparedness to change habits depends not only on their sociodemographic characteristics, but also on their preferences in terms of types of holidays, which is a new result. Still there is a lot of heterogeneity between countries. Given the hierarchical nature of the data, multilevel logistics models are used to take in account both the variability between citizens and the between Member States. The use of such models made it possible to consider also the social context and the specific culture of each country, which influence the willingness of European citizens to adopt sustainable behavior during holidays. This approach identified which countries show greater sensitivity to the issue, as well as a greater propensity of their citizens to change their own habits

Keywords: Sustainability · Tourism · EU · multilevel modeling

1 Introduction

The objective of this paper is to understand which characteristics influence the decisions, among European citizens, to adopt eco-sustainable behavior during their holidays. The data comes from the Flash Eurobarometer 499 survey, carried out in 2021 with the participation of 25,714 European citizens from the 27 Member States (MSs). The questionnaire includes ten questions about participants' intentions to implement various eco-sustainable actions. Citizens' preparedness to change habits depends not only on their sociodemographic characteristics, but also on their preferences in terms of types of holidays, which is a new result. Still there is a lot of heterogeneity between countries. Given the hierarchical nature of the data, multilevel logistics models are used to take in account both the variability between citizens and the variability between MSs. The use of such models made it possible to consider both the differences between Member States, as well as the social context and the specific culture of each country, which influence

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 145–151, 2025.
https://doi.org/10.1007/978-3-031-64431-3_25

the willingness of European citizens to adopt sustainable behavior during holidays. This approach identified which countries show greater sensitivity to the issue, as well as a greater propensity of their citizens to change their own habits.

2 The Database

The data analyzed in this paper were collected by the survey Flash Eurobarometer 499 on attitudes of European citizens towards tourism. Eurobarometer is the polling instrument of the European Union institutions to monitor public opinion and other relevant issues from European citizens and enterprises. Collected data are open and can be downloaded from the portal together with questionnaires and other metadata.

Flash Eurobarometer 499 was conducted in October 2021 and wanted to investigate travel behavior and the impact on it of the Covid-19 pandemic, booking channels and sources of information for travelling arrangements, reasons for selecting destinations, options and information on sustainable tourism. The total sample consists of 25,714 European citizens over 15, distributed in the 27 European Union (EU) MSs. During the interview, conducted with CATI method, questions were asked about behaviors during holidays and travelling, and the impact of the Covid-19 pandemic, the main reasons for choosing the destination and opinions and information on tourism sustainability, including the willingness to change habits to be more sustainable. The survey also collected information on the sociodemographic characteristics of the respondents.

The proportions of male (48%) and female (51%) respondents are approximately similar. The age range covers from 15 to 102 years, mean age is 48.7. Regarding employment status, the majority of participants are employed (47%), while 32% have no professional activity. In relation to family composition, 45% of families are made up of two members over 15 years of age. Before the pandemic, the majority of European citizens used to travel at least once a year, among them, 51% did so several times a year, while 33% did so once or twice. The vast majority of EU citizens expect the COVID-19 pandemic to have some kind of long-term impact on their travel, however, there are still a considerable percentage of people (24%) who do not expect any influence[1]. When organizing trips and holiday activities, European citizens mainly rely on online platforms for professional accommodation services, such as hotels and B&Bs, or on websites that combine travel services, such as accommodation, car rental and flights. However, traditional methods for travel planning remain important, in fact, almost 24% of respondents prefer to organize their own trip through someone they know, while more than 19% continue to use travel agencies. In terms of accommodation preferences, more than 40% of European citizens opt for commercial accommodation with meals included, followed by private accommodation booked directly by the owner, a preferred option by over 32% of those interviewed, and staying in friends' or relatives' places (29.5%). When choosing a destination, European citizens focus on three main aspects: the cultural offer (43.7%), the overall price of the trip (42.9%), and the natural environment (42.6%). Subsequently, with lower percentages, respondents mention as important the activities available in the destination (30.1%) and the accessibility of services and activities for all

[1] All analyses were performed using post-stratification weights in order to guarantee sample representativeness.

(23.9%). The majority of European citizens (56%) said they rely on recommendations from friends, colleagues or family, for information when planning their trip. A good percentage (34.3%) also use websites that collect opinions and ratings from travellers. The second most used source of information, however, is personal experience (37.4%). Promotion of eco-friendly practices is considered as the most important characteristic of a tourist destination by only 14.6% or respondents; and sustainability certification of accommodation and attractions by 14.0% of them.

The analysis of data related to the willingness of European citizens to become aware of sustainability in their travel habits is believed to be a crucial aspect for the development of policies and strategies for sustainable tourism in Europe. This topic was considered in the aforementioned survey and the percentages of positive responses to these questions are shown in Table 1, respondents were asked if prepared to adopt nine specific green actions while talking holidays and, in general, to change their habits while traveling. Table 1 shows, as an important positive fact, that only 14.7% of respondents are not willing to change their travel habits. Furthermore, it is observed that the actions with the highest percentage of affirmative responses are the consumption of local products (55.3%) and the reduction of waste on vacation (48.4%). However, when the adoption of eco-sustainable attitudes requires greater economic effort, the percentage of people willing to implement them decreases: only 32.6% of participants are willing to pay more to benefit the local community, while 35% are willing to do so to protect the natural environment. This result, together with the answers to the important characteristics for a destination, showing low percentages of respondents mentioning eco-friendly practices and sustainability, suggests the need to intervene to encourage the adoption of eco-sustainable behaviors during vacations. At the same time, responses collected in the 27 European countries are quite different.

Table 1. Percentage of YES responses

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

3 Methodology

The multilevel methodology offers several advantages over the traditional regression model that assumes that the data are independent and identically distributed; these are, first of all, the possibility to consider that the variation is both at the individual and group level, in the estimation of the regression coefficients [2]. A second advantage consists in the fact that corrected standard errors are obtained when estimating regression coefficients for particular groups. In the case considered for this analysis, the response variable corresponds to the probability that a European citizen is willing to implement a sustainable action while travelling or on holidays. This is represented by a binary variable $y_{ij} \in \{0,1\}$ that refers to citizen i living in country j and with expected value π_{ij}. The multilevel approach is applied to logistic regression (as well as other generalized linear models) in the same way as for linear regression. The coefficients β_j, that are associated with the explanatory variables, are not constant but vary between the second-level units. It is allowed to include in the regression model covariates from the second level, in order to capture the differences between the second-level units in their influence on the response variable, taking in account the heterogeneity also at this level. In other words, multilevel modelling considers the variation between second-level units that cannot be explained by the model's first-level predictors [1].

4 Results

In order to identify the characteristics of the citizens that influence their predisposition to adopt eco-sustainable behaviors during vacations, different logistic regression models were developed for each one of the ten questions related to what is contained in the questionnaire, as reported in Table 1. The intraclass correlation coefficient (ICC) has been calculated on the observed data for each of the ten response variables, justifying the use of multilevel models. Given that our data has a hierarchical structure and taking in account the differences observed between the countries, multilevel logistic regression models were estimated, using the individual units of analysis, that is, the citizens, as the lower level and the country as the upper level. Three country-level variables were added as explanatory variables to consider the possible influence of specific characteristics of the 27 EU MSs: per capita GDP in 2021, a measure of the importance given in the country to environmental protection, and the bed occupancy rate in October 2021, to represent the impact of the touristic sector in each country; all obtained from the Eurostat website database.

Next, several multilevel logistic models were estimated. Table 2 indicates which is the best fitting model with the mentioned criteria for each action.

Age appears as a statistically significant predictor in all models: older tourists show a disposition to take holidays outside the high season and reduce water consumption while on holidays; younger tourists intend to perform all other actions to behave in a more sustainable way while travelling and are, in general, willing to change their habits to protect the natural environment. All socio-demographic characteristics collected with the survey have an impact at least on some actions and on the preparedness to change. In general, female tourists, employees and those with no activity (mainly retired persons),

components of smaller families, and frequent travelers showed higher probabilities to be prepared to change their travelling and vacation habits to have a more sustainable behavior. In this paper, however, we want to give some attention also to the potential effects of other factors as the main different options and preferences to organize vacations by European citizens, and some specific characteristics of the EU MSs.

Table 2. Best fitting model for each action

Consume locally sourced products on holiday	Random intercept and slope and level-2 covariates
Reduce waste while on holiday	Random intercept and slope and level-2 covariates
Take holidays outside the high tourist season	Random intercept
Travel to less visited destinations	Random intercept
Choose transport options based on ecological impact	Random intercept and slope
Pay more to protect the natural environment	Random intercept and slope and level-2 covariates
Reduce water usage on holiday	Random intercept and slope and level 2 covariates
Contribute to carbon-offsetting activities	Random intercept and slope and level 2 covariates
Pay more to the benefit of the local community	Random intercept
I am not prepared to change my habits	Random intercept and slope and level-2 covariates

These evidences can ben very relevant to design effective green policies. With reference to the preferred type of accommodation, tourists who spend holidays in homes of friends and relatives do not intend to travel to less visited destinations; however, they choose ecological transport and pay attention to consumption of water and waste generation. Citizens choosing camping show a general favorable attitude towards the environment. Choosing a commercial accommodation with board increases the probability to pay more to protect the natural environment and contribute to carbon-offsetting activities; if the accommodation is without board, tourists are in favor of consuming locally-sourced products. Tourist who choose online booking without payment are those more prepared to change their habits.

Tourists using more traditional ways of booking, like the phone or the counter of a travel agency, are less keen to change habits in general, and especially to move to less visited destinations. For those citizens who book their holidays through an online professional platform, the model estimates an opposite result: they show the highest probabilities of performing all sustainability actions except choosing transport based on ecological impact, which is instead an action mentioned by those tourists using online private platforms to book.

Being interested in destinations that received official certifications of sustainability and involve local population in touristic activities increases the probability to be prepared to change habits while travelling and to adopt all nine specific actions proposed in the survey. Information sources used to plan a vacation have a non-negligible impact on a sustainable attitude when travelling: for example, tourists, who refer to travel agencies, are keener to take holidays outside the high tourist season, while those who read specialized magazines are willing to travel to less visited destinations; tourist, who rely on personal experience, are those more prepared to change.

At country level, estimation results show no effect of per capita GDP. Attention to environmental protection has a positive significant effect on some actions: paying more for the benefit of the environment and the local community, and choosing transport options based on ecological impact, and, in general, on the preparedness to change habits to be more sustainable. The percentage of occupancy of beds in hotels has instead a negative effect on the general attitude an on the actions of paying more both for the environment and the community. It appears that in countries where tourism is more important, citizens are less keen to spend money for the scope of a green activity.

5 Conclusions

The results of the analyses presented in this paper showed that only a small percentage (14.7%) of the European population is not willing to change their travel habits, this is an encouraging fact. However, this attitude differs a lot bot within and between European countries. The analyses highlighted that the socio-demographic characteristics of citizens influence their willingness to change their habits while on holiday in order to preserve the environment. A new result emerging in this paper is that citizens' preferences in terms of destination, type of accommodation, sources of information have an impact of their preparedness to change in order to behave in a more sustainable way. Furthermore, the willingness of European citizens to adopt sustainable behavior on holiday is also influenced by the social and cultural context of the country in which they reside, for example citizens living in Spain, Sweden and Slovakia appear the most sensitive to the issue, showing greater willingness, while in countries traditionally associated with tourism such as Italy, France and Croatia there is less preparedness to change habits while travelling; this evidences is supported by the fact that, surprisingly, the importance of the touristic sector in the country showed a significant but negative effect in the models. Per capita GDP di not revealed ad a significant factor.

Policies or interventions aimed at promoting more sustainable behavior during holidays must be adapted or personalized according to the specific characteristics of each country and for different segments of tourists, especially segments that consider difference preferences in organization and type of holidays. It is important to adopt culturally sensitive or targeted approaches to maximize the effectiveness of initiatives to promote more sustainable behaviors. Only a small part of the European population gives priority, when choosing a destination, to the most specific issues of sustainable tourism, such as the promotion of eco-sustainable practices (4.45%), the means of low impact transport (4.15%), sustainability certifications (2.64%) and the involvement of the local population (3.44%). This indicates the need to continue working to raise awareness among European citizens on the topic of tourism sustainability.

References

1. Goldstein, H.: Multilevel Statistical Models. Arnold publisher, London (1999)
2. Hox, J., Moerbeek, M., van de Schoot, R.: Multilevel Analysis. Tecniques and Applications. 3rd edn. Routledge, London (2017)



Extending the Vertical Model: An Alternative Approach to Competing Risks with Clustered Data

Salvatore Battaglia[1(✉)], Marta Fiocco[2], and Hein Putter[3]

[1] Department of Economics, Business, and Statistics, University of Palermo, Palermo, Italy
salvatore.battaglia02@unipa.it
[2] Mathematical Institute, Leiden University, Leiden, Netherlands
[3] Leiden University Medical Center, Leiden University, Leiden, Netherlands

Abstract. Vertical modelling (VM) offers an alternative approach to competing risks models, particularly when relaxing the proportional assumption or when facing missing causes of failure. It focuses on the joint probability, denoted as $P(T, D)$, of the time to failure T and the cause of failure D, which is decomposed into $P(D|T)$ and $P(T)$. Both components rely on observable quantities: the total hazard and the relative cause-specific hazard, which can be easily estimated using multinomial regression and Cox proportional hazard models adding for covariates.

This paper introduces a novel approach by incorporating a random component in each part of the model to address unobserved heterogeneity in the presence of clusters. Data from the EMergenze-URgenze database, comprising Emergency Department (ED) access records from 63 Sicilian EDs in 2019, are utilized to analyze the risk of hospitalization or discharge during the length of stay once admitted to the hospital in a multi-center setting. The Cumulative Incidence Function (CIF) estimated from the vertical mixed model (VMM) is compared with that of the traditional competing risks frailty model.

The VMM with independent random effects represents the first attempt to extend VM and can be viewed as an alternative to the frailty competing risks model when computational time for estimation methods is prohibitive or when exploring different aspects of the data is desired.

Keywords: vertical model · Emergency Department · competing risks · random effects

1 Introduction

Competing risks arise when different types of events can happen, and the occurrence of one event may prevent or influence the occurrence of others.

Within this framework, the focus lies on the joint distribution of time to failure T and the cause of failure D, represented as $P(T, D)$. The typical method for

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 152–157, 2025.
https://doi.org/10.1007/978-3-031-64431-3_26

estimating such a joint probability is through the Cumulative Incidence Function (CIF) by the cause-specific hazard model [1]. However, interpreting CIFs can be challenging due to the lack of a direct relationship between cause-specific hazards and CIFs [2].

In 1999, Fine and Gray introduced the proportional sub-distribution hazard model [3], which acts as a regression model for CIFs. Despite its capability to provide valid inferences, this model relies on a "unsual" risk set construction, where observations experiencing one cause of failure remain at risk for other causes. Over the past two decades, this model has been frequently utilized in medical and healthcare research to predict cumulative incidence in competing risk scenarios, often without full awareness of its limitations [4].

This paper focuses on the vertical model (VM) [5], another approach to address competing risks. The model is based on a decomposition of the joint probability $P(T, D)$ as a mixture model:

$$P(T, D) = P(D|T)P(T) \tag{1}$$

The two components in (1) are observed, namely relative hazard and overall hazard, and can be easily estimated through a multinomial regression model and a Cox proportional hazard model when accounting for covariates.

In this study, we extend the vertical model by introducing random components (U_i, V_i) into both sub-models to accommodate unobserved heterogeneity that may arise in clustered data settings.

2 Definition of the Vertical Mixed Model

We consider, for simplicity, the situation where $K = 2$. To include random effects in the VM (we recommend Nicolaie et.al. [5] for detailed methodological definition), we assume that the clusters are observations coming from a random distribution that must be assumed.

Denote the couple (u_i, v_i) as the random effects for the overall hazard and the relative hazard respectively in the i-th cluster, $B(t)$ a smoothed function of time and Z the covariate vector $(Z_1, Z_2, ..., Z_p)$. Assuming cause 1 to be the reference, the model for the relative hazard is a logistic mixed model, defined as:

$$\pi_2(t) = \frac{exp(\beta_2^T B(t) * Z + u_i)}{1 \ + \ exp(\beta_2^T B(t) * Z + u_i)} \tag{2}$$

where β_2 is the vector of p regression coefficients.

The model for the overall hazard is a Cox frailty model with censorship equal to the observations that do not experience any event:

$$\lambda_\bullet(t) = \lambda_0(t) \ exp(\gamma^T * Z + v_i) \tag{3}$$

where γ is another vector of p regression coefficients.

From (2) and (3) it is worth noting that the couple (u_i, u_i) acts additively in the linear predictors of the two sub-models. Moreover, we assume that the random effects are normally distributed with zero mean and variance-covariance matrix Σ [7]. The random effects represent the center effect for each sub-model and can provide insight into the exit process inside the Hospitals. The correlation coefficient $\rho_{u,v}$ represents the degree of correlation between the two random variables (U, V). In this work we assume independence between (U, V), thus $\rho_{u,v}$=0. The independence assumption is translated by estimating separately the two sub-models.

3 Data

The data we used come from the EMUR database, which is a comprehensive national information system collecting data from Emergency Departments (EDs), including details regarding all accesses to Italian EDs. Specifically, we employed the Sicilian 2019 EMUR dataset consisting of 1,724,758 records from 63 EDs [8].

Each record includes several information. More specifically:

- ED's variables:
  - Hospital's size;
- Socio-demographic variables:
  - Age: the patient's age at the moment of arrival at the ED;
  - Gender.
- Access-related variables:
  - Calendar arrival time (hour, weekday, season), admission time, and leaving time at the ED. LoS is the time of interest used in the analysis and It is computed as the time that occurred since ED's entry to one of the events of interest.
  - Arrival mode: the modality in which patients arrive at the ED (we dichotomized this variable: autonomously or by ambulance);
  - Pathology: after being classified into 11 macro-categories based on clinical criteria using ICD-9-CM codes, we excluded, to make them comparable, accesses coming from *Complications of Pregnancy, Childbirth, and the Puerperium* (ICD-9 [630–679]), *Congenital Anomalies and Certain conditions originating in the perinatal period* (ICD-9 [740–779]), *Neoplasms* (ICD-9 [140–239]), and *Mental Diseases* (ICD-9 [290–389]).
  - Hospitalization/Discharge decision: patients can either be hospitalized or discharged, depending on the physician's decision. These two types of "exits" are our endpoints in the VM.

4 Results

Due to the computational burden, the dataset was reduced to 135.748 accesses from 63 Sicilian EDs. The analysis aimed to determine the probability of being

Hospitalized (or Discharged) during the LoS once the patients are admitted to the Hospital. The model included the pathology, which involved seven different macro-diagnoses, as well as the age and arrival mode of the patients (by ambulance or by themselves) at the ED, as covariates that act as correction factors for pathology.

Since we assumed that the effect of each pathology on the probability of being hospitalized (or discharged) varies during LoS, we introduced an interaction term between a smoothed function of LoS and the pathology in the relative hazards model. Specifically, we chose a cubic spline with 5 knots to capture this interaction, interpolating the relative hazard at the 5 quantiles of LoS (1%, 10%, 50%, 90% and 99%). The selection of these quantiles is based on two factors: firstly, the observation that the great part of events occurred within 1 h until experiencing one of the events, and secondly, to prevent increased computational time in the case of additional knots.

Concerning the model for the relative hazard, accesses involving Diseases of the *Skin and Musculoskeletal system* (*Skin&Musc_Dis*, ICD9 [680–739], OR = 0.08) are least likely to result in hospitalization (and a the same time the most likely to be discharged). Conversely, accesses involving *Disease of the blood* (Blood_Dis, ICD9 [240–289]) are the most likely to be hospitalized (OR = 0.87) Additionally, older patients have a higher probability of hospitalization (1.4% more per year), and accesses arrived through ambulance are more likely to be hospitalized (more than 2.5 times the accesses arrived by own means). The overall hazard coefficient estimates provide complementary information. For example, accesses with skin diseases stay less in the ED (HR = 1.65); on the contrary, accesses with Blood_Dis stay longer in the ED (HR=0.74).

The predicted relative hazard and the overall survival probability have been computed for each pathology (see Fig. 1). The left panel in Fig. 1 shows the probability of hospitalization given admission to the hospital during the LoS, and the right one, either in terms of a hazard or survival scale, shows the intensity of the overall hazard during LoS. This indicates how many accesses remain at risk at a given LoS time and how many have expired.

Examining the pathologies individually, Blood_Dis shows the highest probability of hospitalization within the first 5 h of LoS, subsequently being replaced by *Circulatory and Respiratory Diseases* (Circ&Resp_Dis, ICD9 [350–519]). Both are also associated with longer LoS. In contrast, Inf_Dis shows the lowest hospitalization probability over time.

Figure 2 shows the CIFs computed by the competing risks frailty model according to *Injury and Poisoning* (Inj&Pois, ICD9 [800–999 + E]) and Skin&Musc_Dis. While they may seem comparable, the CIFs of the cause-specific frailty model tend to grow more rapidly in the first hours of LoS than those obtained from the VM, particularly concerning discharges.

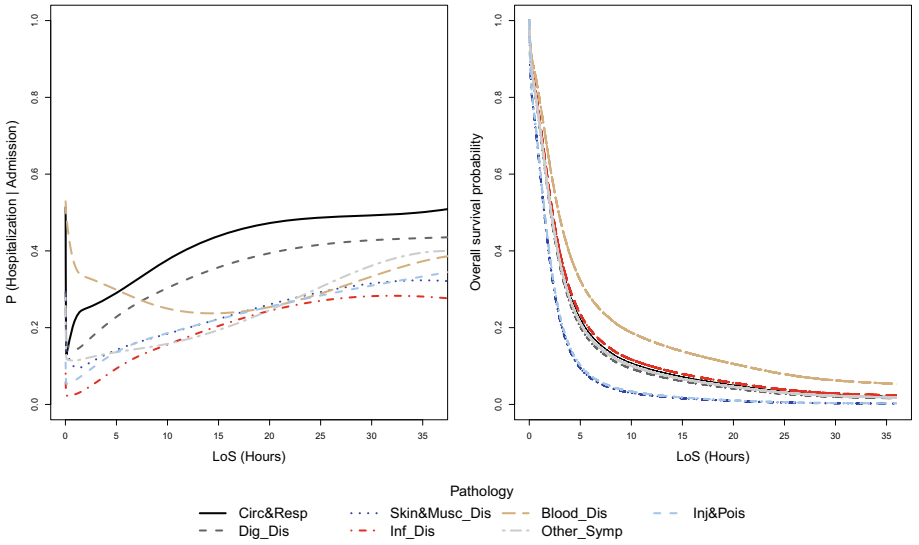


Fig. 1. Predicted probability of being hospitalized given being admitted to Hospital during LoS, or relative hazards (left) and survival probability of being admitted to hospital, or overall hazard (right), according to pathology.

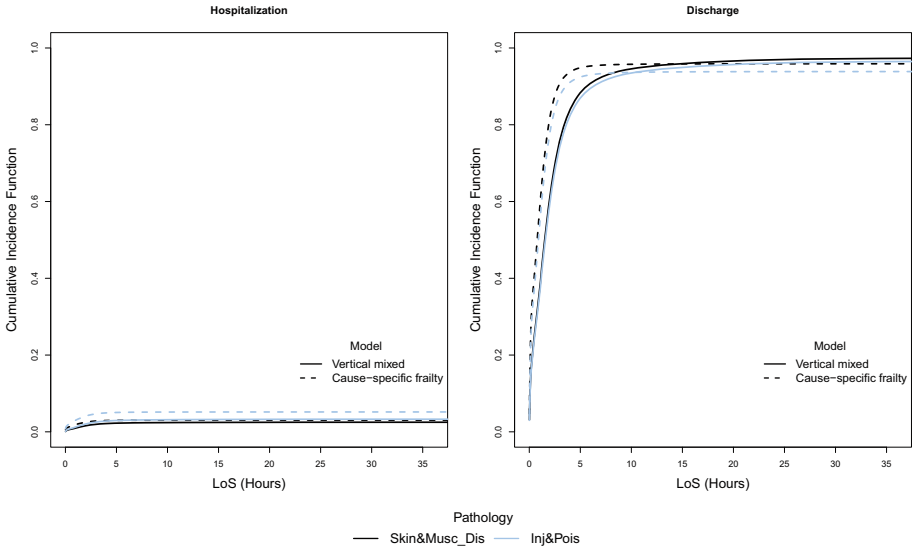


Fig. 2. Comparison between CIFs computed by the VM and the Cause-specific frailty model in accesses with Skin&Musc_Dis and Inj&Pois.

5 Conclusions

We have introduced a novel and alternative approach to conducting competing risks analysis that incorporates a random component to account for unobserved heterogeneity. While the frailty competing risks model, commonly used for this purpose, can be computationally challengingt, he extended computational time required motivated us to explore a faster alternative capable of handling larger datasets. Thus, we developed an extension of the VM, which offers improved computational efficiency and scalability to accommodate large volumes of data.

The interpretation of the coefficient of the VMM and its graphical representation provides insight into the behavior of the relative and overall hazard at specific time points. For example, accesses with Blood_Dis have the longest LoS and the highest rates of hospitalization up to 5 h, indicating they may present the most challenging cases in the decision-making process.

References

1. Prentice, R.L., Kalbfleisch, J D., Peterson Jr, A.V., Flournoy, N., Farewell, V.T., reslow, N.E.: The analysis of failure times in the presence of competing risks. Biometrics, 541–554 (1978)
2. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. Stat. Med. **26**(11), 2389–2430 (2007)
3. Fine, J.P., Gray, R.J.: A proportional hazards model for the subdistribution of a competing risk. J. Am. Stat. Assoc. **94**(446), 496–509 (1999)
4. Austin, P.C., Fine, J.P.: Practical recommendations for reporting Fine-G ray model analyses for competing risk data. Stat. Med. **36**(27), 4391–4400 (2017)
5. Nicolaie, M.A., van Houwelingen, H.C., Putter, H.: Vertical modeling: a pattern mixture approach for competing risks modeling. Stat. Med. **29**(11), 1190–1205 (2010)
6. Rueten-Budde, A.J., Putter, H., Fiocco, M.: Investigating hospital heterogeneity with a competing risks frailty model. Stat. Med. **38**(2), 269–288 (2019)
7. Christian, N.J., Ha, I.D., Jeong, J.H.: Hierarchical likelihood inference on clustered competing risks data. Stat. Med. **35**(2), 251–267 (2016)
8. Siciliana, R.: LINEE DI INDIRIZZO PER LA GESTIONE DEL SOVRAFFOLLAMENTO NELLE STRUTTURE DI PRONTO SOCCORSO DELLA REGIONE SICILIA.



Likelihood-Based Boosting for Variance Components Selection in Linear Mixed Models

Michela Battauz(✉) and Paolo Vidoni

Department of Economics and Statistics, University of Udine, Udine, Italy
michela.battauz@uniud.it

Abstract. A novel boosting method for the selection of the variance components in linear mixed models have been recently proposed. The method relies on the directions of negative curvature to deal with a non-convex loss function, which is the negative profile log-likelihood. Differently from previous proposals in the literature, which focus on the selection of the predictors in the fixed part of the model, the procedure provides a selection of the random part. More specifically, starting from a model without random effects, the algorithm sequentially includes them by updating the parameters that determine their covariance matrix. After summarizing the proposal, this paper presents an original application.

Keywords: negative curvature direction · nonconvex optimization · random effects

1 Introduction

Boosting is a statistical method that allows for fitting a model and performing variable selection and regularization at the same time. Some proposals in the literature developed boosting algorithms for random effects models. Generalized linear mixed models were considered in [6], [7] focused on binary and ordinal random effects models, while [8] dealt with on additive mixed models. These proposals focus on the selection of the variables in the fixed part of the model, while the variables in the random part should be pre-specified. To the best of our knowledge, the only work that considered the selection of the variables with a random effect is given by [5]. However, the random effects are considered only in the supplementary material of that paper and a study of the performance of the proposal using real or simulated data is missing. Recently, [2] proposed a boosting method to select the random effects of a linear mixed model. The purpose is to provide a automatic selection of the random part of the model, which is accomplished exploiting an optimization method based on the directions of negative curvature [4] to deal with the non-convexity of the objective function to minimize.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 158–163, 2025.
https://doi.org/10.1007/978-3-031-64431-3_27

This paper is structured as follows. The procedure is briefly reviewed in Sect. 2, an original application is presented in Sect. 3, while Sect. 4 contains some concluding remarks.

2 Model and Method

Consider the following linear mixed model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \widetilde{\mathbf{Z}}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i,$$

where i is the index of the clusters, $i = 1, \ldots, m$, $\mathbf{y}_i$ is a vector that includes the response values, $\mathbf{X}_i$ is a matrix of predictors, $\widetilde{\mathbf{Z}}_i = \mathbf{Z}_i\boldsymbol{\Delta}^\top$, $\mathbf{Z}_i$ is a matrix of predictors obtained as a subset of $\mathbf{X}_i$, the random effects $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\boldsymbol{\Delta}^\top\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Omega})$, since $\boldsymbol{\Delta}^\top\boldsymbol{\Delta} = \boldsymbol{\Omega}$, and the errors $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Furthermore, we adopt the decomposition $\boldsymbol{\Delta}^\top = \boldsymbol{\Lambda}\boldsymbol{\Gamma}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix and $\boldsymbol{\Gamma}$ is a lower triangular matrix with ones on the diagonal. Hence, the non-zero variances of the random effects are determined by the elements of $\boldsymbol{\Lambda}$ only. Let $\boldsymbol{\theta}$ be the vector that includes the parameters of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ and $\ell_p(\boldsymbol{\theta})$ the profile log-likelihood function which can be derived along the lines of [1].

The boosting method proposed in [2] starts from $\boldsymbol{\theta} = \mathbf{0}$ and, at each step, updates the parameters according to either a Newton direction or a negative curvature direction, depending on which provides the largest reduction of a quadratic approximation of the objective function to minimize, that is $-\ell_p(\boldsymbol{\theta})$. Since the aim is to provide a sparse solution, the procedure follows a component-wise approach, considering only one parameter at a time and selecting the component that leads to the largest gain. Let $\widehat{g}_c^{(m-1)}$ and $\widehat{H}_c^{(m-1)}$ be the gradient and the Hessian of component c of $\boldsymbol{\theta}$ computed at step $m-1$. The Newton direction is given by

$$s_c^{(m)} = -\widehat{g}_c^{(m-1)} / \widehat{H}_c^{(m-1)}$$

while the negative curvature direction is

$$d_c^{(m)} = -sign(\widehat{g}_c^{(m-1)})\widehat{u}_c^{(m-1)},$$

where $\widehat{u}_c^{(m-1)}$ is equal to 1 if $\widehat{H}_c^{(m-1)} < 0$, and it is null if there is not a negative curvature direction, which means that all the elements on the diagonal of the Hessian matrix are not negative. After the selection of the direction that leads to the largest reduction of a quadratic approximation of the objective function among all Newton and negative curvature directions, the corresponding element c^* of $\boldsymbol{\theta}$ is updated. If the selected direction is a Newton one, the component c^* is updated as follows:

$$\widehat{\theta}_{c^*}^{(m)} = \widehat{\theta}_{c^*}^{(m-1)} + \nu s_c^{(m)};$$

otherwise, the new value of the component c^* is given by:

$$\widehat{\theta}_{c^*}^{(m)} = \widehat{\theta}_{c^*}^{(m-1)} + \gamma d_c^{(m)}$$

where ν and γ are small step lengths. The simulation studies reported in [2] show the good performance of the method in terms of correct selection of the variables with random effects and of prediction accuracy attained.

3 A Real-Data Application

Our proposal was applied to the analysis of lipid changes in middle-age women using the data from the Study of Women's Health Across the Nation (SWAN), a multi-site, longitudinal, epidemiological study designed to examine the health of women during their middle years. The data are freely available from the website https://www.icpsr.umich.edu/web/ICPSR/series/00253. Analogously to [3], we excluded women taking lipid lowering medications at all visits and, considering only subjects without missing data at all the visits, we obtained a dataset composed of 7 observations over time for 3093 subjects. The response variable we considered is high density lipoprotein (HDL) cholesterol and the covariates are selected on the basis of previous studies [3]. The numerical covariates that vary over time, and hence possibly having a random effect, are age, BMI, estradiol and follicle-stimulating hormone (FSH). The other covariates include ethnicity, smoking status at baseline, concurrent diabetes at baseline, percent of kilocalories from fat and percent kilocalories from alcoholic beverages both measured at baseline through a questionnaire, time spent walking or biking at baseline, time spent doing sport or exercise at baseline, taking blood pressures medications, taking cholesterol medications, which are both dummy variables that could vary over time. As in [3], we used baseline smoking status because this variable presents few changes over time. We used the baseline values for the time spent walking or biking and doing sport since they were not recorded in two visits. Both these variables are on an ordinal scale. All numerical variables were standardized. The boosting algorithm used $\nu = 0.1$ and $\gamma = 0.1$. Figure 1 shows the estimates of the parameters in $\boldsymbol{\theta}$ and the objective function as a function of the iterations. In order to reduce the randomness involved in cross-validation, we repeated the procedure 20 times and used the average cross-validation error to select the number of iterations where to stop.

As a result, we selected 479 iterations, which determined the following covariance matrix of the random effects:

$$\begin{pmatrix} 0.5543 & 0.0185 & -0.0053 & 0 & 0.0198 \\ 0.0185 & 0.0228 & -0.0002 & 0 & 0.0007 \\ -0.0053 & -0.0002 & 0.0063 & 0 & -0.0002 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0198 & 0.0007 & -0.0002 & 0 & 0.0079 \end{pmatrix}.$$

Hence, the procedure selected a random effect for all the variables but one. More specifically, the first one refers to the random intercept and the others to age, BMI and FSH, while estradiol does not have a random effect. The covariances between these random effects are all different from zero. The random intercept presents the larger variability, while the other variances are quite small. This should be considered together with the fixed effects, which present small values too (0.073 for age, -0.379 for BMI and 0.121 for FSH), and the estimate of the variance of the error, namely $\hat{\sigma}^2 = 0.157$. The complete fixed-effects estimates are reported in Table 1.

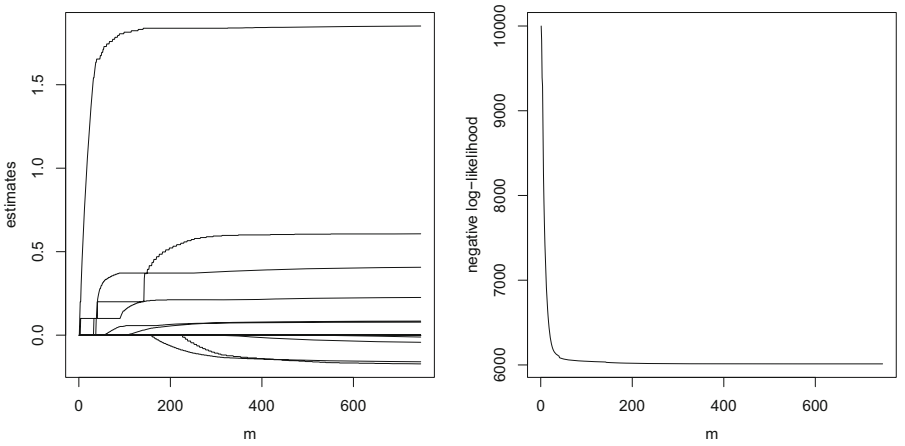


Fig. 1. The boosting method in the real-data example: on the left panel estimates of the parameters over the iterations; on the right panel the objective function over the iterations.

Table 1. Fixed effects estimates in the SWAN study.

variable	levels (for categorical variables)	estimates
Intercept		0.242
Age		0.073
BMI		−0.379
estradiol		0.091
FSH		0.121
Ethnicity (base = Black/African American)	Chinese/Chinese American	−0.129
	Japanese/Japanese American	−0.291
	Caucasian/White Non-Hispanic	−0.310
Smoking (base = Never smoked)	Past smoker	0.055
	Current smoker	−0.366
Diabete	Yes	−0.197
percent of kilocalories from fat		0.083
percent of kilocalories from alcoholic beverages		0.193
blood pressures medications	Yes	−0.053
cholesterol medications	Yes	0.002
walking or biking (base = Never/< 5 Minutes Per Day)	5–15 Minutes/Day	0.002
	16–30 Minutes/Day	0.009
	31–45 Minutes/Day	0.130
	More Than 45 min/Day	0.222
Sports (base = Never)	Less Than Once A Month	−0.047
	Once A Month	0.011
	2–3 Times A Month	−0.162
	Once A Week	−0.032
	More Than Once A Week	0.056

Besides using the boosting method, we also fitted various models by MLE using the R package lmm. They are reported in Table 2, together with the log-likelihood, AIC and BIC. To avoid trying all combinations, we adopted a hierarchical procedure. Starting from the random intercept model, we added one random effect for all the potential variables in turn and selected the one that leads to the highest log-likelihood. This random effect was then kept into the model and the random effects for the other variables were added in turn, till the full model was reached. Looking at AIC, the selected model includes random effects for all the variables, while BIC leads to a more parsimonious solution with three random effects. Note that, using the hierarchical approach, we fitted 11 models, instead of $4^2 = 16$ combinations of variables (keeping always the intercept into the model). Testing also all possible combinations of variances and covariances equal to zero or not results in $14^2 = 196$ models to fit, which becomes quite cumbersome. Instead, the boosting method performs an automatic selection of both variances and covariances.

Table 2. Models fitted by MLE in the Swan study. The * indicates the highest log-likelihood for each number (#) of parameters, the lowest AIC and BIC among all the models.

variables with random effects	# parameters	log-likelihood	AIC	BIC
Intercept	26	−6114.4	12281	12462
Intercept, age	28	−6046.4 *	12149	12344
Intercept, BMI	28	−6094.2	12244	12440
Intercept, estradiol	28	−6103.0	12262	12457
Intercept, FSH	28	−6072.0	12200	12395
Intercept, age, BMI	31	−6026.4 *	12115	12331 *
Intercept, age, estradiol	31	−6043.6	12149	12365
Intercept, age, FSH	31	−6032.8	12128	12344
Intercept, age, BMI, estradiol	35	−6023.5	12117	12361
Intercept, age, BMI, FSH	35	−6012.5 *	12095	12339
Intercept, age, BMI, estradiol, FSH	40	−6002.3	12085 *	12364

4 Conclusions

The boosting procedure proposed in [2] revealed effective in automatically selecting the random effects in the application presented in this paper. Although the solution found might not seem very parsimonious, it should be noted that the AIC would lead to include a random effect for all the variables. On the basis of BIC, only three variables would have a random effect. However, such criterion did not provide a good performance in the simulation study presented in [2].

The proposal seems promising for dealing with other models, especially in case of non-convex objective functions to minimize.

References

1. Bates, D.M., DebRoy, S.: Linear mixed models and penalized least squares. J. Multivar. Anal. **91**, 1–17 (2004). https://doi.org/10.1016/j.jmva.2004.04.013
2. Battauz, M., Vidoni, P.: A boosting method to select the random effects in linear mixed models. Biometrics **80**, 1–9 (2024). https://doi.org/10.1093/biomtc/ujae010
3. Derby, C., Crawford, S., Pasternak, R., Sowers, M., Sternfeld, B., Matthews, K.: Lipid changes during the menopause transition in relation to age and weight. Am. J. Epidemiol. **169**, 1352–1361 (2009). https://doi.org/10.1093/aje/kwp043
4. Gould, N.I.M., Lucidi, S., Roma, M., Toint, P.L.: Exploiting negative curvature directions in linesearch methods for unconstrained optimization. Optim. Method Softw. **14**, 75–98 (2000). https://doi.org/10.1080/10556780008805794
5. Kneib, T., Hothorn, T., Tutz, G.: Variable selection and model choice in geoadditive regression models. Biometrics **65**, 626–634 (2009). https://doi.org/10.1111/j.1541-0420.2008.01112.x
6. Tutz, G., Groll, A.: Generalized linear mixed models based on boosting. In: Kneib, T., Tutz, G. (eds.) Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir, pp 197–215. Springer-Verlang, Berlin (2010)
7. Tutz, G., Groll, A.: Likelihood-based boosting in binary and ordinal random effects models. J. Comput. Graph. Stat. **22**, 356–378 (2013). https://doi.org/10.1080/10618600.2012.694769
8. Tutz, G., Reithinger, F.: A boosting approach to flexible semiparametric mixed models. Stat. Med. **26**, 2872–2900 (2007). https://doi.org/10.1002/sim.2738



Financial Institution Confidence Index with Natural Language Processing

Leonardo Becarelli and Luca Trapin(✉)

Alma Mater Studiorum Università di Bologna, 40126 Bologna, Italy
{leonardo.becarelli2,luca.trapin}@unibo.it
https://www.unibo.it/sitoweb/leonardo.becarelli2/en,
https://www.unibo.it/sitoweb/luca.trapin/en

Abstract. This work applies natural language processing (NLP) methods to analyze Sect. 7 (Management Discussion and Analysis) of the 10-K filings for a large set of U.S. financial companies in the period 2005–2022 in order to build an index of *financial institution confidence*. We perform sentiment analysis using the *FinBERT* pre-trained financial language representation model for financial text mining, and aggregate the yearly sentiment values of each financial institution to build our index. We show a positive correlation between our index and the main events that characterize the US economic cycle.

Keywords: BERT · financial institutions · sentiment analysis · 10-K filings

1 Motivation

Corporate filings serve as hubs of information, offering deep insights into a company's financial health and strategies, and are the foundation onto which most of the computational analysis builds to compute performance measures and build reports [4]. Central to financial analysis are the Securities and Exchange Commission 3 (SEC) filings, which serve as primary sources of information for publicly traded companies in the United States. Of particular importance is the Form 10-K, an annual report filed in compliance with the generally accepted accounting principles (GAAP) by publicly traded companies at the end of each fiscal year, that thoroughly documents the company's business performance, including details such as company history, organizational structure, and a breakdown of its financial status over the entire year. To pursue this research, Sect. 7 *"Management Discussion and Analysis"* of the 10-K report has been analyzed. This section is comprised of a cohesive description made by company analysts of financial results, giving an extensive overview of the company's economic dynamics through the eyes of the company's management. This gives stakeholders a peek at the financial cycles that regulate the company's operations and the main fundamental reasons behind how the business has been operated through the fiscal

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 164–169, 2025.
https://doi.org/10.1007/978-3-031-64431-3_28

year, allowing for the description and analysis of what could impact the company in the foreseeable future. This forward-looking aspect of Sect. 7 is key to understanding its importance as a tool for financial analysis.

NLP can be applied to extract and interpret the sentiments contained in these expansive texts. One example of architectures used for this type of analysis is the Bidirectional Encoder Representations from Transformers (BERT). This is an open-source machine-learning framework developed by researchers at Google in 2018 and can be defined as the "state-of-the-art NLP model" at the time of writing. The revolutionary aspect comes from the introduction of a technology called "transformer", which allows for unprecedented scalability of computation and to include in the training process large amounts of un-labeled corpus, a feat impossible before its introduction. The improvement that BERT has brought to financial statement analysis is undeniable. For example, [5] noticed a "significant enhancement of applying BERT in financial sentiment analysis when compared with the existing models".

The purpose of this paper is to show how 10-K reports of financial institutions can be used in combination with the BERT architecture to perform sentiment analysis and build an index of financial institution confidence by aggregating yearly sentiment scores. This index provides indication on the confidence of financial institutions on the future development of the U.S. financial ecosystem and it can be used as a proxy of financial stability.

2 Data Retrieval and Preprocessing

The U.S. Securities and Exchange Commission allows users to freely access its reports using the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR). Through various features, the EDGAR API[1] can be used to perform searches, fetch specific portions of documents, and much more. The tools employed in this research are the following:

- the Query API, allowing the search and filter (of) all 18+ million filings and exhibits published on SEC EDGAR since 1993. The API accepts simple and complex filter expressions and returns filings in JSON[2] format;
- the Full-Text Search API, a tool allowing to search the full text of all EDGAR filings submitted since 2001 and downloading of predefined portions of each filing.

The first step in gathering the documents was defining the documents that will be used in the analysis and fetching their Uniform Resource Locator (URL), i.e. their position on the web, through the Query API. The Query API supports the main types of search parameters, such as the type of filing, the ticker of

[1] An application programming interface (API) is an interface enabling data transmission between two softwares.

[2] JSON (JavaScript Object Notation) is a text format for storing and transporting data. This type of file provides a human-readable format for storing and manipulating data when developers build software.

the company, the date and time of filing, and many more. To build the query request, we have made use of 3 parameters:

1. the *filedAt* parameter, allowing to restrict the search for filings filed within the years of interest (2005–2022).
2. the *formType* parameter, used to bring up only 10-K filings among all the available document types;
3. the *entities.sic* parameter, to restrict the search to companies only related to certain industries making use of Standard Industrial Classification (SIC) codes[3]

The Query API returns a JSON file containing the name of the company, the date the documents were filed, and a list of every document and appendix comprised in the filing, with their corresponding URL. The URL's were then fed to the Full-Text Search API for the retrieval of the actual text. Through text mining, we specifically selected the information regarding the 10-K filing for every year of interest. To download the corpus of text we have then specified a new class called *SECExtractor*. This class handles the construction of the query request for the Full-Text Search API and the process required to automate this process. In this second query request, two additional parameters were added:

1. the *section_id* parameter, used to extract only Sect. 7, Management Discussion and Analysis, from the report;
2. the *output_type*, to force the extractor to output the section in text format for easier analysis.

The resulting JSON file contains the corpus of the "Management Discussion and Analysis" section for the filings used in the analysis. After ensuring the JSON files presented complete and exhaustive corpora, we proceeded by preparing the documents for the sentiment analysis procedure. As the model requires text to be segmented into sentences to build the tokens, the function *sent_tokenize* from the *Natural Language ToolKit* package of Python was used. With each corpus segmented into sentences, we proceeded with the main focus of the analysis, i.e. the implementation of the sentiment analysis model using NLP. In particular, we use a special model built on the BERT architecture called *finBERT*, see [2]. This is a pre-trained NLP model specifically built to tackle financial texts.

3 Sentiment Analysis Using BERT

The implementation of the finBERT model can be summed up as a classification problem. For this analysis, we have mainly leveraged two Python libraries:

– *PyTorch*, an "optimized tensor library for deep learning using GPUs and CPUs". Here it is used mostly to allow for the use of tensors and to combine the computational power of CPU and GPU cores to streamline the process;

[3] We use the codes of financial-related industry sectors (6021, 6035, 6011, 6200, 6311).

– *Transformers*, used to load pre-trained models and a tokenizer for text processing. In this analysis, the model in question is finBERT.

For each sentence of the corpus, finBERT returns the probability that the sentence belongs to each of three possible sentiment classes: positive (1), neutral (0), and negative (-1). We then assign each sentence to the class with highest probability, and build a sentiment score for the entire corpus, with a score ranging from -1 to 1, by averaging across the classes associated to the sentences forming the corpus.

4 Financial Institutions Confidence Index

We build sentiment scores for $K = 10,000$ randomly selected financial institutions for each year from 2005 to 2022. We consider as financial institutions all the companies belonging to one of the following SIC code: national commercial banks (6021), savings institution federally chartered (6035), federal credit agencies (6011), security and commodity brokers, dealers, exchanges and services (6200), life insurance (6311). To compute our financial institution confidence (FIC) index, we aggregate the yearly sentiment scores of each institution for each year t. In particular, the index is calculated as follows:

$$FIC_t = \frac{1}{K} \sum_{k=1}^{K} (ss_{k,t} - \overline{ss}_t)^2$$

where $ss_{k,t}$ is the sentiment score for the kth institution in the tth year, $\overline{ss}_t$ is the mean sentiment score of all institutions in the tth year. We claim this index can be used as a measure of financial uncertainty in the spirit of [1]. The rationale is as follows: when institutions share the same view and a common opinion on the future of financial markets, the index takes small values, signalling stability in the system; in contrast, when perceptions of the institutions differ, the index raises signalling more uncertainty about the future. As shown in [3], sentiment analysis methods have the ability to capture at least parts of the CCI using information non-directly related to consumer confidence like Twitter data. As aggregate yearly sentiment scores are computed using SEC filings, documents produced directly by agents from financial institutions and vetted by third-party regulators, the resulting index should be a good proxy for institutional confidence in the economy.

Figure 1 depicts the dynamic of the FIC index for each year from 2005 to 2022. We can note the index is high before a period of financial instability occurs, confirming that our index is informative of the future stability of the financial system. In particular, we can see that the index peaks right before the Global Financial Crisis, a period of turmoil that financial analyst started to perceive back in 2005 with the beginning of the difficulties encountered by the U.S. housing market. In contrast, we can see that the index did not peak before the COVID-19 period, despite the market collapse, because the event was

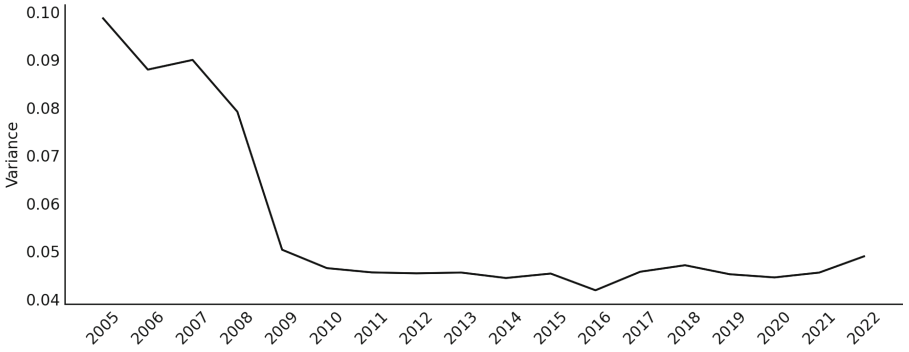


Fig. 1. Financial institutions confidence index

completely exogenous. The behaviour of our index is coherent with that of some components of the *OFR Financial Stress Index*, especially the *credit*, *safe assets*, *equity evaluation* components.

5 Conclusion

Despite being a relatively recent field, the use of sentiment analysis to capture elusive metrics has been one of the more studied topics in recent years, with extensive use in publications from all sectors. Financial and economic studies have been at the forefront of this revolution due to the intrinsic "emotional" nature of some of its agents and how they behave in the environment may not always be measurable by common means; this is where sentiment analysis can really be pivotal. The rapid development in open-source deep learning architectures like BERT has made powerful instruments available to the general public to a scale previously unimaginable, massively increasing the reliability of models thanks to their accessibility and ease of computational scalability. The results produced are promising, showing a link to the various economic cycles of the United States, with particular attention to systematic shocks in the financial world like the 2008 financial crisis. This correlation can be used as additional proof that sentiment analysis, developed using more and more sophisticated methods, will be key in future economic and financial research. The resulting values produced, computed by analyzing documents produced by financial institutions, could be used as a new way to measure how these institutions feel about the future state of the economy and to capture sentiments that may be hidden by personal and company interests.

References

1. Altig, D., Barrero, J.M., Bloom, N., Davis, S.J., Meyer, B., Parker, N.: Surveying business uncertainty. J. Econ. **231**(1), 282–303 (2022)

2. Araci, D.: Finbert: financial sentiment analysis with pre-trained language models (2019). URL https://arxiv.org/abs/1908.10063
3. Colladon, A.F., Grippa, F., Guardabascio, B., Costante, G., Ravazzolo, F.: Forecasting consumer confidence through semantic network analysis of online news. Sci. Reports **13** (2023). https://doi.org/10.1038/s41598-023-38400-6. URLhttps://www.nature.com/articles/s41598-023-38400-6
4. Pankoff, L.D., Virgil, R.L.: On the usefulness of financial statement information: A suggested research approach. Acc. Rev. **45**(2) (1970). URLhttps://www.jstor.org/stable/244378
5. Zoen, J., Huang, X., Mou, H., Li, D., Wu, Q., Xu, Y.: Bert-based financial sentiment index and lstm-based stock return predictability (2019). URLhttps://arxiv.org/abs/1906.09024



Nonparametric Space-Time Density Estimation for Point Processes over Irregular Regions

Blerta Begu[1(✉)], Simone Panzeri[2], Eleonora Arnone[3], Michelle Carey[1], and Laura M. Sangalli[2]

[1] School of Mathematics and Statistics, University College Dublin, Belfield Dublin 4, Dublin, Ireland
blerta.begu@ucdconnect.ie

[2] MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milan, Italy

[3] Dipartimento di Management, Università degli Studi di Torino, Corso Unione Sovietica 218 bis, 10134 Turin, Italy

Abstract. In this work, we propose a nonparametric density estimation technique for space-time inhomogeneous Poisson point processes. We employ a penalized likelihood framework able to handle event data occurring over spatial regions with complex shape. The regularization term, guided by partial differential equations, ensures smoothness in the estimate. To substantiate our method, we provide theoretical validation. For the estimation procedure, we rely on advanced numerical techniques. Moreover, we incorporate uncertainty quantification tools into our methodology. Finally, we demonstrate the effectiveness of our proposed approach through simulation studies and an application to epidemiological data.

Keywords: Space-time density estimation · Poisson point process · partial differential equation regularization · nonparametric approach

1 Introduction

The growing availability of geolocated event data, continuously recorded across both space and time, has attracted considerable interest in the literature, with a particular focus on the development of methods for density and intensity estimation. In Fig. 1, we provide a visual representation of a real case scenario, illustrating the temporal evolution of gastrointestinal infections in Hampshire county, UK; see, e.g., [18].

Space-time point processes have been extensively explored in the literature and find applications across various domains, as detailed in [6,7]. These applications range from disease surveillance and crime detection to distance sampling in biomarine populations and monitoring occurrences of wildfires and hurricanes.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 170–175, 2025.
https://doi.org/10.1007/978-3-031-64431-3_29

Seismology, that is another significant field of interest, has been extensively investigated using ETAS models, as discussed in [13].

Many existing methods simplify the analysis by aggregating time, primarily due to computational constraints, as observed in approaches such as Space-Time Kernel Density Estimation and Log-Gaussian Cox Process, outlined in [19,20]. Other methods handle space and time separately, attempting to discern spatio-temporal interactions afterward. Unfortunately, this often leads to the loss of critical information inherent in the non-separability of the spatio-temporal process. Additionally, the involvement of multiple variables complicates solution interpretation, as highlighted in [9,10]. Furthermore, only a few existing methods effectively handle irregular regions. Notably, one such method is the hierarchical model-based Bayesian Integrated Nested Laplace Approximations (INLA) with stochastic partial differential equations, as introduced in [12].

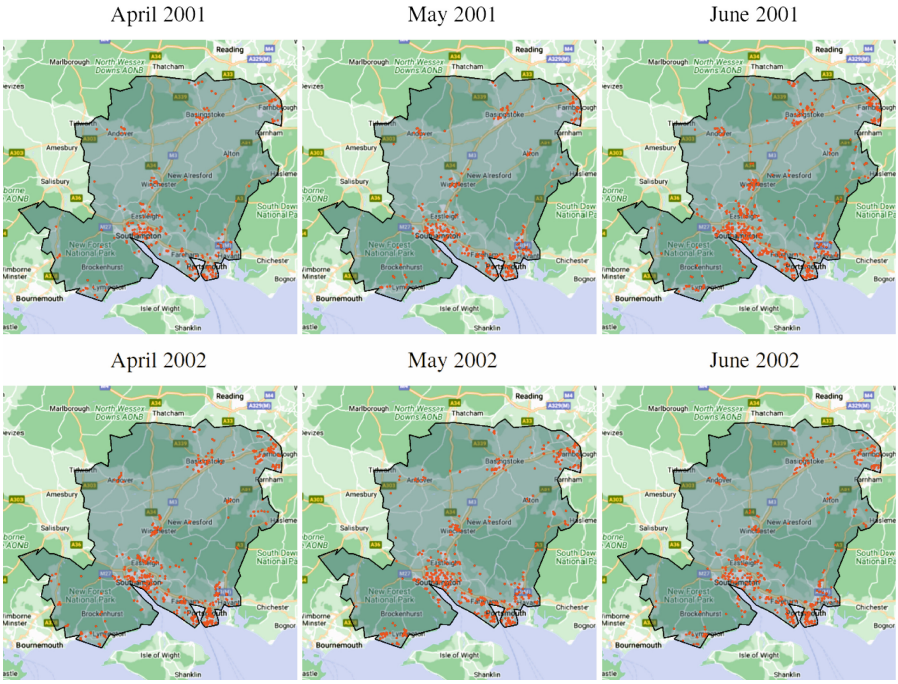


Fig. 1. Disease surveillance data. Locations of gastrointestinal infections in Hampshire county, UK. Data recorded from 1st October 2001 to 31st March 2002. The point pattern highlights that the phenomenon is expressed differently along the two shores of the Southampton Bay.

Given the limitations of some of the existing approaches, in this work, we introduce a novel method for space-time density estimation. The method is able to accommodate observations that are continuous over space and time, and does

not impose any restriction on the separability of space and time dimensions. Additionally, it is able to handle non-convex spatial regions, making it a decisive and efficient tool for analyzing complex spatio-temporal phenomena.

We employ a nonparametric approach based on maximum likelihood estimation with roughness penalties. Specifically, we design appropriate regularizing terms by leveraging partial differential operators in both space and time. This enables us to capture signals exhibiting intricate and multi-modal behaviors in the spatio-temporal domain.

This research has contributed to the enlargement of a class of nonparametric methods able of handling functional and spatial data across complex domains, through regularization techniques involving partial differential equations (PDEs), as discussed in [16]. This class of methods has so far been used in spatio-temporal domains only under regression settings, see, e.g., [1,5]. The proposed approach extends a simpler model for density estimation over spatial domains only, that has been previously developed for planar and curved domains by [2,8].

Full details of our method can be found in [4].

2 Space-Time Density Estimation

Let us consider a compact subset $\mathcal{D}$ of Euclidean space $\mathbb{R}^2$ and a bounded interval $T = [T_1, T_2]$ within $\mathbb{R}^+$. The spatio-temporal domain of interest is defined as $\mathcal{D}_T = \mathcal{D} \times T$. Let $\{(\mathbf{p}_i, t_i)\}_{i=1}^n \in \mathcal{D}_T$ denote n independent and identically distributed realizations drawn from a probability distribution F with density function $f : \mathcal{D}_T \to \mathbb{R}^+$. Our objective is to find a reliable estimate of f over the domain $\mathcal{D}_T$ via maximum likelihood estimation. It is essential to ensure that f satisfies the following standard properties: (i) $f \geq 0$ in $\mathcal{D}_T$ (ii) $\int_{\mathcal{D}_T} f = 1$. To fulfill requirement (i), we estimate the log-density $g = \log f$ instead of f. Meanwhile, requirement (ii) is directly satisfied within a proper formulation of the objective functional. Specifically, our statistical model involves minimizing the following penalized functional:

$$\mathcal{L}(g) := -\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{p}_i, t_i) + \int_{\mathcal{D}_T} e^{g(\mathbf{p},t)}\, d\mathbf{p}\, dt + R(g; \lambda_{\mathcal{D}}, \lambda_T). \tag{1}$$

The functional in (1) comprises three terms: the negative log-likelihood as the first term, a second term ensuring requirement (ii), and finally the regularizing term $R(g; \lambda_{\mathcal{D}}, \lambda_T)$ with smoothing parameters $\lambda_{\mathcal{D}}, \lambda_T > 0$. We employ a penalizing term that separately manages the regularity of the function in space and in time. Specifically, it is designed as follows:

$$R(g; \lambda_{\mathcal{D}}, \lambda_T) := \lambda_{\mathcal{D}} \int_{\mathcal{D}_T} \left(\Delta\, g(\mathbf{p}, t)\right)^2 d\mathbf{p}\, dt + \lambda_T \int_{\mathcal{D}_T} \left(\frac{\partial^2 g(\mathbf{p}, t)}{\partial t^2}\right)^2 d\mathbf{p}\, dt,$$

where Δ is the Laplace operator. The desired estimate $\hat{g}$ of g is then obtained by solving an unconstrained minimization problem.

We investigate the main theoretical properties, such as the uniqueness and existence of a minimizer for the functional $\mathcal{L}(g)$, as well as the consistency of the estimator $\hat{g}$. These results are closely tied to the selection of the penalizing term. We formalize these properties by introducing a suitable functional space and meticulously extending the reasoning from simpler univariate contexts, as stated by [11,17], to space-time domains.

Since it is not possible to obtain a closed analytical expression for the estimator, we address the problem by solving it with advanced numerical approximations. In particular, we use a suitable space-time separable basis expansion for the spatio-temporal field by employing finite elements in space and cubic B-spline basis in time. These choices allow us to work with spatial regions having non-convex shapes and to achieve effective computational performances. Subsequently, efficient optimization algorithms and cross-validation techniques are employed to determine the optimal smoothing parameters and obtain the desired estimate.

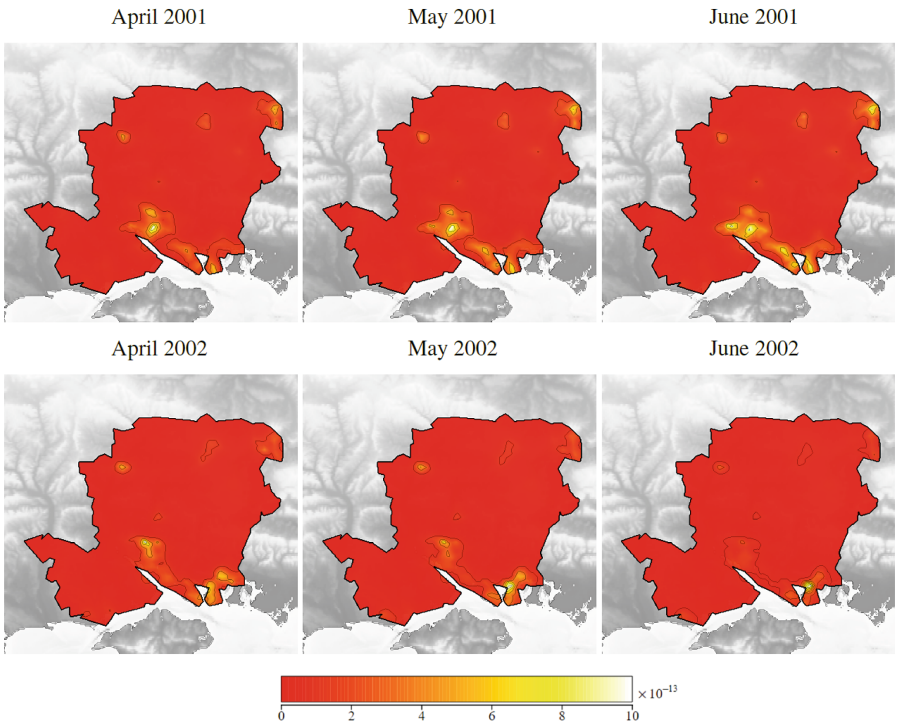


Fig. 2. Application to disease surveillance data: density estimates of gastrointestinal virus occurrences in Hampshire County from October 1st, 2001, to March 31st, 2002. The proposed method accurately handles point patterns over non-convex spatial regions. The estimate correctly captures the different expressions of the phenomenon under study along the two shores of the Southampton Bay.

Finally, confidence bands for the density estimate are constructed by using a Bayesian approach, firstly proposed by [14] in the simpler univariate context. The presented method is implemented in the `R` library `fdaPDE` [3].

3 Application to Disease Surveillance Data

Figure 2 showcases the versatility of our method by applying it to a real-world spatio-temporal point process. Specifically, we analyze the data concerning gastrointestinal infections in the Southampton area, Hampshire, UK, which can be accessed from the supplementary material of [18]. One strength of our method lies in its capacity to handle spatial domains with complex shapes, including those with concavities, that might occur due to complex topography, as illustrated in the case of Hampshire county in Fig. 1. Here the Southampton Bay creates a strong concavity in the domain, and influences the phenomenon under study, as apparent from the point patter in Fig. 1. As highlighted by Fig. 2, the proposed method returns accurate estimates of the spatio-temporal density, being able to correctly capture the different expressions of the phenomenon under study along the two shores of the Southampton Bay.

Acknowledgements. B. Begu acknowledges funding by NexSys SFI/21/SPP/ 3756, Ireland. L.M. Sangalli acknowledges funding by the European Union - NextGenerationEU, research project 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

1. Arnone, E., Azzimonti, L., Nobile, F., Sangalli, L.M.: Modeling spatially dependent functional data via regression with differential regularization. J. Multivariate Anal. **170**, 275–295 (2019)
2. Arnone, E., Ferraccioli, F., Pigolotti, C., Sangalli, L.M.: A roughness penalty approach to estimate densities over two-dimensional manifolds. Comput. Stat. Data Anal. **174**, 107527 (2022)
3. 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
4. Begu, B., Panzeri, S., Arnone, E., Carey, M., Sangalli, L.M.: A nonparametric penalized likelihood approach to density estimation of space-time point patterns. In: Spatial Statistics (2024). https://doi.org/10.1016/j.spasta.2024.100824
5. 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. In: Stochastic Environmental Research and Risk Assessment, vol. 31, no. 1, pp. 23–38 (2017)
6. Daley, D.J., Vere-Jones, D., et al.: An introduction to the theory of point processes: volume I: elementary theory and methods. In: Springer (2003)
7. Diggle, P.J., Moraga, P., Rowlingson, B., Taylor, B.M.: Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. In: Statistical Science (2013)

8. Ferraccioli, F., Arnone, E., Finos, L., Ramsay, J., Sangalli, L.: Nonparametric density estimation over complicated domains. J. Roy. Stat. Soc. Series B, Stat. Meth. **83**(2), 346–368 (2021)
9. Gabriel, E., Rowlingson, B.S., Diggle, P.J.: stpp: an R package for plotting, simulating and analyzing Spatio-Temporal Point Patterns. J. Stat. Soft. **53** 1–29 (2013)
10. González, J.A., et al.: Analysis of tornado reports through replicated spatiotemporal point patterns. J. Roy. Stat. Soc. Series C, Appl. Stat. **69**(1), 3–23 (2020)
11. Gu, C., Qiu, C.: Smoothing spline density estimation: Theory. In: The Annals of Statistics, pp. 217–234 (1993)
12. 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. Stat. Soc. Series B, Stat. Meth. **73**(4), 423–498 (2011)
13. Ogata, Y.: Space-time point-process models for earthquake occurrences. In: Annals of the Institute of Statistical Mathematics, vol. 50, no. 2, pp. 379–402 (1998)
14. O'Sullivan, F.: Fast computation of fully automated log-density and log-hazard estimators. SIAM J. Sci. Stat. Comput. **9**(2), 363–379 (1988)
15. Ramsay, T.: Spline smoothing over difficult regions. J. Roy. Stat. Soc. Series B (Stat. Methodol.) **64**(2) 307–319 (2002)
16. Sangalli, L.M.: Spatial regression with partial differential equation regularisation. Int. Stat. Rev. **89**(3), 505–531 (2021)
17. Silverman, B.W.: On the estimation of a probability density function by the maximum penalized likelihood method. In: The Annals of Statistics, pp. 795–810 (1982)
18. Taylor, B.M., Davies, T.M., Rowlingson, B.S., Diggle, P.J.: lgcp: an R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. J. Stat. Softw. **52** 1–40 (2013). https://www.jstatsoft.org/article/view/v052i04
19. Taylor, B.M., Davies, T.M., Rowlingson, B.S., Diggle, P.J.: lgcp: Log-Gaussian Cox Process (2022). https://cran.r-project.org/web/packages/lgcp/index.html
20. Zhang, Z.: stam: Spatio-Temporal Analysis and Modelling (2010). https://cran.r-project.org/contrib/main/00Archive/stam/



Generalized Additive Mixed Models in Medicine: A Case Study on LDL Cholesterol in People Living with HIV Under Different Antiretroviral Regimens

Michele Bellomo[1(✉)], Nicola Gianotti[1], Camilla Muccini[1], Hamid Hasson[1], Diana Canetti[1], Silvia Nozza[1,2], Monica Guffanti[1], Emanuela Messina[1], Vincenzo Spagnuolo[1], Riccardo Lolatto[1], Sara Diotallevi[1], Antonella Castagna[1,2], and Laura Galli[1]

[1] IRCCS San Raffaele Scientific Institute, Infectious Diseases Unit, Milan, Italy
bellomo.michele@hsr.it
[2] Vita-Salute San Raffaele University, Milan, Italy

Abstract. Generalized Additive Models (GAMs) are statistical methods that enable the modeling of variables through smooth effects. GAMs are powerful and flexible models, well-consolidated among statisticians. However, their application in medicine, especially in infectious diseases, is still limited. In this study, we applied GAMs to evaluate the dynamic of low-density lipoprotein (LDL) cholesterol in people living with HIV (PLHIV) who have switched to an antiretroviral regimen based on doravirine, rilpivirine, dolutegravir or bictegravir. Data were collected at IRCCS San Raffaele Scientific Institute and involve 2742 individuals and 54552 observations. A Generalized Additive Mixed Model (GAMM) and a quasi-Poisson GAMM were trained considering several risk factors and accounting for temporal dynamics through random effects and autoregressive covariance structures. Doravirine exhibited the most favorable effect on LDL cholesterol, with an associated average reduction in the first two years of use of −5.43 mg/L per year according to the GAMM and of −5.41% per year in the quasi-Poisson GAMM.

Keywords: Generalized Additive Mixed Models · mgcv · HIV · HDL cholesterol

1 Background

Generalized Additive Models (GAMs) are statistical methods that generalize linear models by allowing the nonparametric modeling of variables and estimating their effect on the response variable through a generic smooth function. This relaxes the assumption of linearity and captures nonlinear relationships, while maintaining good model interpretability.

GAMs were introduced in the 1980 s [1] and have become a well-known and established methodology among the statistical community. Over time, several

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 176–181, 2025.
https://doi.org/10.1007/978-3-031-64431-3_30

libraries and packages have been developed for their convenient use [2]. Nevertheless, their application in the medical field, particularly in infectious diseases, remains rare.

In this study, we compared the dynamic of low-density lipoprotein (LDL) cholesterol in people living with HIV (PLHIV) who have switched to an antiretroviral regimen based on doravirine (DOR), rilpivirine (RPV), dolutegravir (DTG) or bictegravir (BIC). Dyslipidemia is a common issue in PLHIV, caused by multiple factors and, sometimes, worsened by side effects of antiretroviral drugs. Dyslipidemia is a clinical condition that leads to an increased risk of cardiovascular events and thus higher mortality. Understanding the effects of different antiretroviral drugs on this condition is crucial for physicians to choose the best therapy and improve the quality of life of PLHIV.

2 Methods

2.1 Study Design and Cohort

Observational cohort study on data collected in a large database during routine visits of PLHIV followed at the Infectious Diseases Unit of the IRCCS San Raffaele Scientific Institute (Milan, Italy). The study inclusion criteria are: adult PLHIV, antiretroviral experienced, who switched to a antiretroviral regimen containing DOR, RPV, DTG, or BIC, naïve to the studied drugs at the therapy switch, with at least 2 LDL cholesterol determinations before the switch and at least 4 determinations after the switch. Follow-up accrued from the first LDL cholesterol determination within a 5 year interval before the switch, until the last available determination after the switch or the start of an antiretroviral regimen not including any of the considered drugs (DOR, RPV, DTG, BIC). The same individual may take more than one drug under study during follow-up, either at different times, due to a subsequent change in therapy, or simultaneously (an antiretroviral therapy is generally constituted by a mix of antiretroviral drugs).

2.2 Covariates

In addition to DOR/RPV/DTG/BIC intake, the following variables were considered: age, sex, body mass index (BMI), presence of comorbidities (diabetes, cardiovascular events), CD4+ cell count (type of white blood cell essential for the immune system and targeted by HIV, whose count indicates immune health in PLHIV), viremia (count of HIV copies present in a millimeter of blood, crucial for assessing the level of viral suppression in PLHIV), use of statins (drugs used to lower cholesterol levels in the blood), use of antiretroviral drugs other than DOR/RPV/DTG/BIC, grouped by class (NRTI, NNRTI, PI, INSTI, boosters), with the exception of tenofovir disoproxil fumarate (TDF) and tenofovir alafenamide (TAF), that were considered separately. TDF has a well-known lipids-lowering effect, while the role of TAF on lipids is still debated. Since TDF and TAF are often given in combination with the study drugs and in a mutually exclusive manner, taking their influence into account is essential to remove possible confounding factors.

2.3 Statistical Modeling

Generalized Additive Mixed Models (GAMM) were used to study the LDL cholesterol dynamic. Covariates related to antiretroviral drugs and statins were coded as time-dependent variables, counting the number of years of continuous use. Continuous variables were modeled non-parametrically via smooth effect, except for the study's target drugs, which were modeled parametrically via β regression coefficient to facilitate comparisons.

Individual characteristics and temporal dynamics were modeled using random effects (random intercept + random slope on the years from the first switch) and an autoregressive covariance structure (first-order AR(1)).

Among the four studied antiretroviral drugs, some are more recent than others. For example, RPV was put on the market in 2011, while DOR in 2018. As a consequence, follow-up duration greatly varies across the four drugs, and therefore we decided to estimate slopes and confidence intervals related to these drugs distinguishing two intervals: the first 2 years of continuous use and longer periods (2 years is the longest period for which there were sufficient data to estimate the effect of the drug with shorter follow-up).

Finally, we trained a quasi-Poisson GAMM to deal with the overdispersion of GAMM residuals. A quasi-Poisson is a model trained using quasi-likelihood, a quantity with properties similar to likelihood, but which only requires assumptions on the mean-variance relationship for its definition. In a quasi-Poisson model, the variance of the response variable y_i has the form $V(\mu) = \phi\mu$, where V is the variance function, μ the mean of y_i and $\phi > 0$ a proportionality factor. The quasi-Poisson model can be seen as a generalization of the Poisson model (in the Poisson model the variance and the mean of y_i are equal, case $\phi = 1$). The ϕ parameter indicates the degree of dispersion of the quasi-Poisson compared to the standard Poisson, and the case $\phi > 1$ is suitable for count data that are more dispersed than a Poisson distribution. Since no specific assumptions are made on the distribution, the quasi-Poisson is appropriate also for continuous data. In the quasi-Poisson model we used the default log-link function, and therefore estimates are interpreted in terms of percentage changes.

All the models were fitted using R version 4.3.0 and the `mgcv` package [3].

2.4 Results

Overall, 2742 participants met the inclusion criteria, for a total of 54552 determinations.

All smooth effects are significant at the 5% level, except for NRTI other than TDF and TAF, INSTI other than BIC and DTG, and PI. Figure 1 shows the estimated smooth effects of the significant variables.

Estimated effects have non-linear behaviors, which justifies the use of GAMs and their non-parametric modeling. LDL cholesterol appears to increase with age up to about 60 years, and then start to decrease. The LDL cholesterol is higher in people with higher BMI till 35 kg/m^2, where the effect of BMI reaches a plateau (for BMI values greater than 35 kg/m^2, the estimated smooth effect has higher

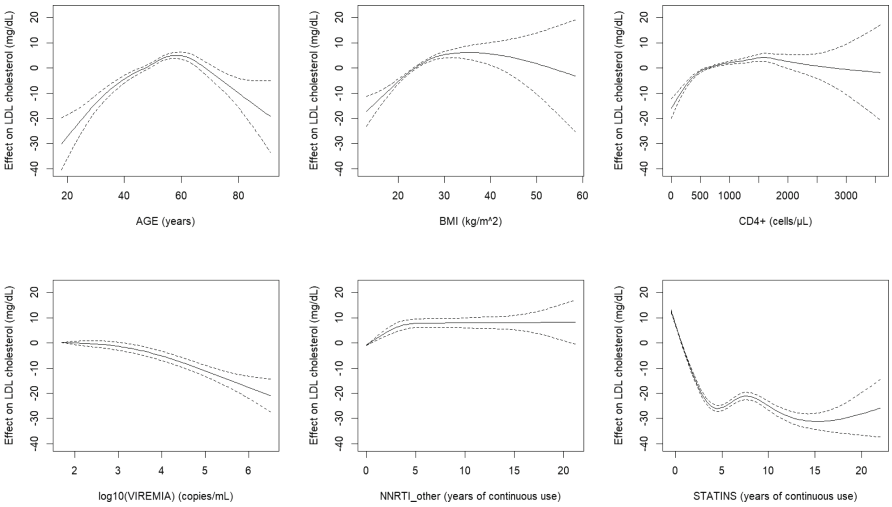


Fig. 1. Estimated significant smooth effects

margins of uncertainty due to the small number of individuals with such BMI values). LDL cholesterol decreases for CD4+<500 cells/μL and viremia>10^3 copies/mL, indicating that people more immunocompromised and with active infection have lower cholesterol. Statins have an important lowering effect, which peaks after 5 years of use and settles around a total average decrease of 25 mg/dL. We interpret the oscillations around the steady-state value as noise. Boosters (not shown in Fig. 1) and NNRTI other than DOR and RPV have similar increasing effects, which peak after 4–5 years of continuous use and settle around a total average increase of 10 mg/dL.

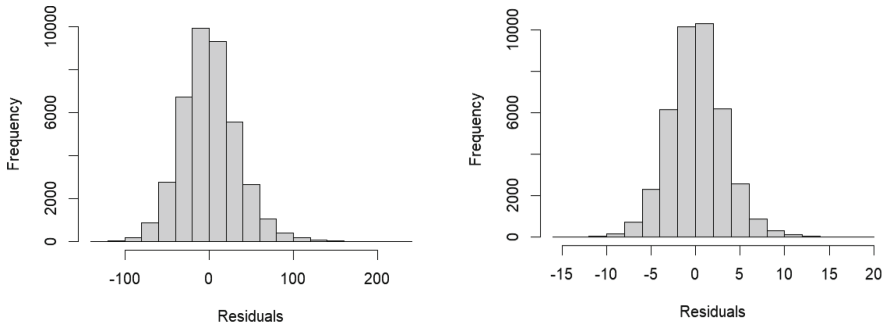
Concerning parametric modeled variables, the estimated intercept is 119.75 mg/dL (p<0.001). There are no significant differences in LDL cholesterol between men and women (mean difference 0.17 mg/dL, p=0.905). A diabetes diagnosis leads to a mean decrease in LDL cholesterol of −4.97 mg/dL (p=0.002), while the occurrence of a cardiovascular event leads to a mean decrease of −14.00 mg/dL (p<0.001). These effects can be justified by healthier lifestyles and introduction of permanent lipid-lowering therapies following these diagnoses. The lipid-reducing effect of TDF is confirmed in the first 2 years of use (mean reduction of -3.77 mg/dL per year, p<0.001), while for periods > 2 the effect changes (average increase of 0.44 mg/dL per year, p<0.001). Regarding TAF, a slight decreasing effect is detected for periods longer than 2 years (average decrease of -1.28 mg/dL per year, p=0.004), while there are no significant effects for the first 2 years of use (average increase of 0.28 mg/dL per year, p=0.569). β coefficients, 95% confidence intervals and p-values related to the drugs under study are reported in Table 1. As the median follow-up for DOR was 1.2 years (IQR 0.6–1.7), no slope could be estimated for the time interval > 2 years.

RPV, DTG, and BIC have statistically significant but modest effects on LDL cholesterol in the first 2 years of use. DOR, instead, has an important reduction effect of −5.43 mg/L per year in the first 2 years of use. For periods greater than 2 years, a statistically significant effect is detected only for RPV (slight increase of 0.69 mg/L per year).

Checking the model assumptions, we notice some overdispersion in the right tail of the residuals. We tackle this problem by training a quasi-Poisson GAMM. In this way, the overdispersion is eliminated and the distribution of deviance residuals is symmetric, as shown in Fig. 2. Shapes of smooth effects, estimated coefficients and p-values obtained in the quasi-Poisson GAMM are consistent with the GAMM with Gaussian distribution. In particular, DOR is associated with a mean decreasing effect of −5.41% per year in the first two years of use (p=0.003).

Table 1. β estimates, 95% confidence intervals and p-values for the mean annual changes in LDL cholesterol related to the drugs under study

	Mean change per year 0–2 years	Mean change per year >2 years
DOR	–5.43 mg/L [–9.78, –1.08] p = 0.014	NA
RPV	1.01 mg/L [0.26, 1.76] p = 0.008	0.69 mg/L [0.26, 1.12] p = 0.002
DTG	1.10 mg/L [0.43, 1.77] p = 0.001	–0.31 mg/L [–0.76, 0.15] p = 0.184
BIC	–1.59 mg/L [–2.90, –0.27] p = 0.018	1.11 mg/L [–1.15, 3.37] p = 0.335



(a) GAMM (b) quasi-Poisson GAMM

Fig. 2. Histogram of residuals for model diagnostic

3 Conclusions

We applied GAMs to study LDL cholesterol in PLHIV under different antiretroviral regimens. Smooth effects enabled a more detailed modeling of variables and a consequent better estimate of errors and variances. At the same time, by modeling the study's target drugs parametrically via β coefficient, we obtained the main results in a format familiar to physicians and more suitable for comparisons and dissemination. The flexibility of the `mgcv` library enabled the inclusion in the models of random effects, covariance structures and distributions alternative to the Gaussian. The quasi-Poisson GAMM showed a better fit to the data thanks to a less dispersion of the residuals. However, we presented the results of both models to demonstrate the robustness of our findings and, acknowledging the differing interpretations of estimated coefficients due to the distinct link functions, to offer a more comprehensive and general presentation of the results.

Among the drugs targeted in the study, DOR showed the best outcome on LDL cholesterol, with an estimated average reduction in the first two years of use of -5.43 mg/L per year in the GAMM and of -5.41% per year in the quasi-Poisson GAMM.

We believe this work can motivate more researchers to explore GAMs, essential instruments in the toolkit of a modern biostatistician.

Acknowledgment. Supported in part by a research grant from Investigator-Initiated Studies Program of Merck Sharp & Dohme LLC. The opinions expressed in this paper are those of the authors and do not necessarily represent those of Merck Sharp & Dohme LLC.

References

1. Hastie, T., Tibshirani, R.: Generalized additive models. Stat. Sci. **1**(3), 297–310 (1986)
2. Wood, S.N.: Generalized additive models: an introduction with R. CRC press (2017)
3. Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. Roy. Stat. Soc. Series B: Stat. Methodol. **73**(1), 3–36 (2011)



An Age-Adjusted Disproportionality Measure for Pediatric Pharmacovigilance

Pietro Belloni[1(✉)] and Federica Pizzolato[2]

[1] Department of Statistical Sciences, University of Padua, Padua, Italy
pietro.belloni@unipd.it
[2] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padua, Padua, Italy

Abstract. Adverse Drug Effects (ADEs) pose significant risks, especially in vulnerable populations such as children, who often lack representation in clinical trials. In this work, we explored the identification of statistically significant associations between drug administration and ADEs (pharmacovigilance signals) in pediatric patients using spontaneous drug safety databases. By applying the Regression adjusted gamma-Poisson shrinkage (RGPS) model, which produces age-adjusted disproportionality measures, the study aims to improve the precision and accuracy of signal detection compared to other models. The results suggest that the RGPS model selects signals with greater precision, particularly when considering age effects. This emphasizes the importance of age-adjusted models in pharmacovigilance efforts to enhance drug safety monitoring and regulatory decision making for pediatric populations.

Keywords: Drug safety · Disproportionality analysis · Pediatrics pharmacovigilance · Age adjustment

1 Introduction

Adverse Drug Effects (ADEs) pose a significant threat, contributing to hospitalizations, deaths and substantial financial burdens on public healthcare systems [1]. During the development of a new drug, pharmaceutical companies are required to conduct clinical trials to assess its safety and detect ADEs. However, clinical trials do not allow the identification of all possible ADEs, some of which are not observed since they occur with such frequency that observation during the clinical trial phase is unlikely. For this reason, a post-marketing surveillance phase is necessary to continuously monitor the safety of marketed drugs. In this phase, the use of spontaneous adverse drug effects reporting databases is fundamental in order to analyze potential associations between a drug and an adverse event. If such an association is found, it can lead to the withdrawal of the drug from the market (or restrictions on its use).

The impact of ADEs on special populations such as children (0–21 years) is particularly concerning, as they often lack representation in clinical trials evaluating the safety and effectiveness of drugs under development [2]. As a result,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 182–188, 2025.
https://doi.org/10.1007/978-3-031-64431-3_31

only a limited number of drugs are available for children and drugs originally developed for adult patients are often given to pediatric patients. This may cause adverse reactions in pediatric patients with severe and longer-lasting consequences: evidence suggests that children are more susceptible to ADEs that lead to hospitalizations compared to adults [3]. Each year, millions of children receive prescribed medications and experience ADEs, however pediatric pharmacovigilance is a branch of pharmacovigilance that is still little investigated with fewer data available. Notably, children are characterized by physiological and metabolic changes during growth, such as enzyme activity, that differ from one age group to another and can significantly alter how they respond to a treatment. Therefore, it is important to consider age of pediatric patients when performing a pharmacovigilance analysis in the pediatric setting. Most methods used in literature do not consider age-adjusted disproportionality measures.

Our goal is to identify, through the analysis of children-related data in spontaneous pharmacovigilance databases, the statistically significant associations between the administration of a drug and the occurrence of an ADE in children. In order to better identify associations between drugs and adverse events in children, we compare the age-adjusted disproportionality measure produced by a recent algorithm called RGPS (*Regression-adjusted gamma-Poisson shrinkage*), with other methods currently used by pharmacovigilance authorities (often not age-adjusted).

2 Materials and Methods

2.1 FAERS Data and KidSIDES Data

To ensure post-marketing drug surveillance, pharmacovigilance agencies maintain databases of spontaneous reporting of drug adverse events. These databases collect reports that spontaneously communicate an event after the administration of a drug. These reports can be filed by anyone (patients, physicians, pharmacists, researchers, etc.). Pharmacovigilance agencies constantly scan spontaneous databases to detect signals, that is, statistically significant associations between drugs and ADEs. For a more in-depth discussion of spontaneous databases, their use and limitations, see [4].

Among the spontaneous databases, the *FDA Adverse Event Reporting System* (FAERS) is maintained by the U.S. Food and Drug Administration. Unlike other spontaneous pharmacovigilance databases, FAERS is publicly available. Due to this important feature, FAERS has proven to be a fundamental tool for monitoring the safety of marketed drugs and developing statistical techniques useful for pharmacovigilance.

Since we are interested in using FAERS data restricted to pediatric age reports, we used the KidSIDES database. KidSIDES is an open source database derived from FAERS and pre-processed to select pediatric reports [5]. To obtain a more accurate analysis, we adjusted the association between drugs and ADEs

according to the age of the child. The age was first standardized and then classified into the seven stages of child development according to the *Eunice Kennedy Shriver National Institute of Child and Human Development* [7].

2.2 Regression-Adjusted Gamma-Poisson Shrinkage Model

To scan spontaneous databases, pharmacovigilance authorities make use of many ad-hoc statistical models, both frequentist and Bayesian, often called disproportionality models [6]. The Regression-adjusted gamma-Poisson Shrinkage model (RGPS) [8] is a recent empirical Bayesian disproportionality model that combines the gamma-Poisson Shrinkage model (GPS) [9] and Logistic Regression (LR) for pharmacovigilance data [10]. By combining these two methods, RGPS aims to achieve the best of both. Like LR, RGPS is designed to account for confounders caused by concomitant drugs and drug masking effects. Like GPS, the shrinkage induced by the model protects from false positive drug–ADE associations. Furthermore, unlike GPS, RGPS is able to compute estimates for cases where the number of events reported for a given drug is zero. Since RGPS processes one response (adverse event) at a time, the computation of each response is independent of other responses and the shrinkage model is estimated separately for each response.

The RGPS model consists of three main steps. First, a set of predictors is automatically selected to fit a LR model. In our case, the set of predictors consists of dummy variables indicating the possible administration of selected drugs and the age classes. The selection of the drug used in the LR model is based solely on the number of events (N_j) reported for a given drug j. As proposed by [8], to select drugs, we use a GPS with a $Gamma(\delta_0, \delta_0)$ prior distribution. The resulting LR provides an estimated probability of a response event being in a report depending on the patient's age and the set of drugs in that report. However, the coefficients of this model are merely used to adjust for concomitant drugs as described in the second step.

In the second step, expected counts for each drug in the database (E_j) are estimated based on the results obtained in the LR model computed in the first step. We have that

$$\mathrm{E}_j = \sum_{i=1}^{N} \frac{X_{ij}}{1 + \exp\left(-\beta_{i0} - \sum_{k=1}^{p} \beta_k X_{ik} + \beta_j X_{ij}\right)},$$

where N is the number of unique subject in the KidSIDES dataset and $X_{ij} = 1$ if the potential drug predictor j is included in the report i, $X_{ij} = 0$ otherwise. β_{i0} is the coefficient referring to the age group of the i-th subject and β_j are the coefficients of the LR model, if a drug j is not included in the LR model $\beta_j = 0$.

In the third step, a two-parameter GPS model is used to compute adjusted relative reporting rates and their associated confidence intervals based on observed counts N_j and expected counts E_j, this GPS adopts a single gamma prior with parameters (γ, δ) rather than a mixture of two gammas. The posterior

mean of a drug relative reporting ratio, known as the *empirical Bayes relative report ratio* (EBRRR), is

$$\text{EBRRR}_j = \frac{N_j + \gamma}{E_j + \delta}$$

and the lower and upper bounds of the confidence interval are computed using the posterior distribution $Gamma(N_j + \gamma, E_j + \delta)$. If the 5^{th} percentile of the posterior distribution is higher than 1, the data support the hypothesis of an association between the drug and the ADE involved with a confidence of 90%.

Since RGPS is not available in any public software package, we wrote the R code according to DuMouchel's original formulation and used it to fit the RGPS model with the KidSIDES data.

3 Results

We evaluated the detection performance of RGPS and compared it with the performance of other algorithms using the reference database developed by the *Global Research in Pediatrics* (GRiP) *consortium* [11]. First, we select from KidSIDES a set of drug–ADE pairs with epidemiological or mechanistic evidence in children observed in GRiP ($N = 74$). Then, for each unique drug–ADE pair included in this reduced database, we compute five disproportionality measures: ROR, PRR, age-adjusted ROR (RORadj), GPS and RGPS. Notice that PRR, ROR and GPS measures do not take age into account in any way. To identify the differences between these measures, we examine the point estimate and its associated 90% confidence interval. A pair is considered a signal when the interval lower bound is greater than 1. In the interest of brevity, we have reported only a few drug–ADE pairs, but the conclusions derived from the results obtained apply to all examined pairs.

We noticed that RGPS selects, in most of the ADEs examined, a lower number of significant drug–ADE pairs (Fig. 1.A). The number of signals detected for the four different adverse events reported in Fig. 1.A by ROR, PRR and RORadj is much higher than the percentage of signals detected by GPS and RGPS. The latter two models show similar results, even though RGPS is more conservative and allows for a lower number of false positives. Furthermore, we notice that RGPS has the highest accuracy, measured using the association observed using GRiP as the gold standard (Fig. 1.B).

Figure 2.A shows the results for the drug–ADE pair *Ibuprofen* and *anaphylactic reaction*. Estimates of the ROR, PRR, RORadj and GPS are higher than RGPS values, both considering the point estimate and the confidence interval. In most of the 22 significant drug–ADE pairs, RGPS produces more conservative (smaller) disproportionality measures and tighter confidence intervals than the other models. Furthermore, in Fig. 2.A a smaller value of RGPS suggests an effect induced by the age of the child not captured by the other models. The age effect is noticeable in Fig. 2.B: the two indices adjusted by age are larger than the others. However, RGPS has the upper limit of the confidence interval smaller than RORadj and, for *suicide* adverse drug effect, RGPS selects a lower

number of significant pairs (Fig. 1.A). In some cases, RGPS identifies signals not detected by the other models. For example, RGPS identifies the association between the drug *Ibuprofen* and the adverse event *rash papular* (Fig. 2.C), this pair is listed as a verified positive association in the GRiP reference set. In other cases, a drug–ADE pair is detected from the other measures, but not from RGPS. For example, the association between *Valproic acid* and the *paranoia* adverse drug effect is not in the GRiP reference set and there is no evidence of association in the medical literature, however, ROR, PRR, RORadj and GPS found a significant association (Fig. 2.D).

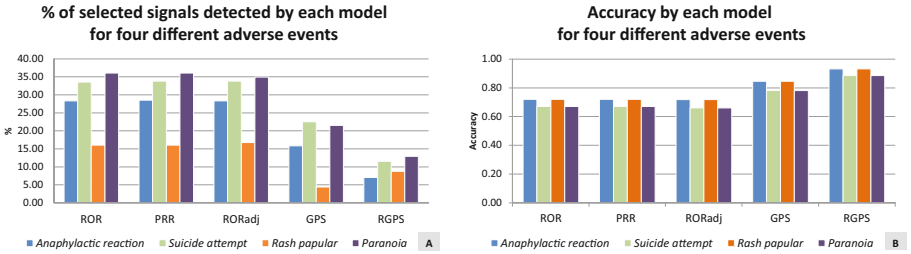


Fig. 1. Percentage of selected signals detected by the models ROR, PRR, RORadjusted, GPS and RGPS (A) and values of accuracy (B) for the ADE: *Anaphylactic reaction*, *Suicide attempt*, *Rash papular* and *Paranoia*.

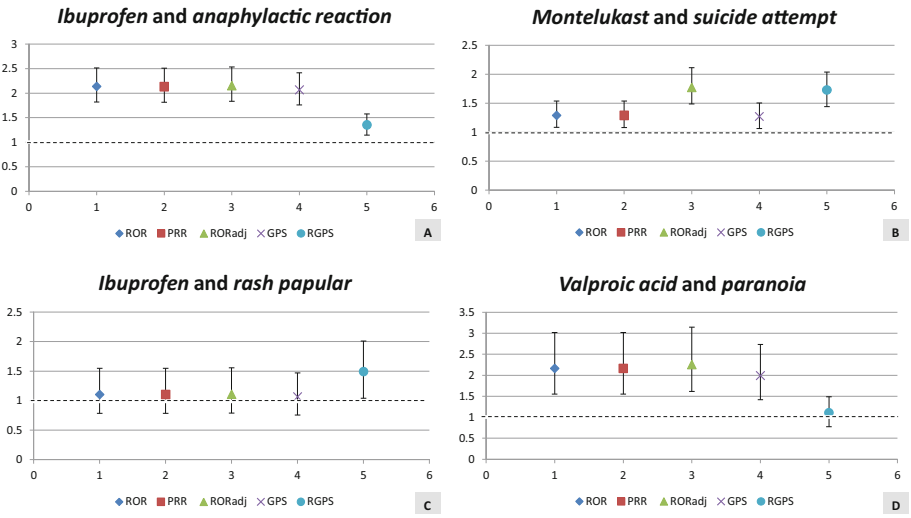


Fig. 2. The point estimates and their associated 90% CI for *Ibuprofen–Anaphylactic reaction* (A), *Montelukast–Suicide attempt* (B), *Ibuprofen–Rash papular* (C) and *Valproic acid–Paranoia* (D).

4 Conclusion

We applied the Regression adjusted gamma-Poisson shrinkage (RGPS) model to produce an age-adjusted disproportionality measure to better identify associations between drugs and adverse events in children. The RGPS model allows estimates that are both shrinked and adjusted by age. Overall, our results suggest that this model produces a more conservative (smaller) disproportionality measure than the GPS and tighter confidence intervals than the other models. Moreover, RGPS selects signals with more precision if we compare its performance with competitors using the GRiP reference set.

Since the RGPS estimates are corrected by age, it is possible to identify some drug–ADE pairs whose association is masked by the age effect and other pairs where the association is significant only if age is also taken into account. In conclusion, this model proves to be of paramount importance for spontaneous pharmacovigilance database scanning procedures.

References

1. Lazarou, J., Pomeranz, B.H., Corey, P.N.: Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. J. American Med. Assoc. **279**(15), 1200–1205 (1998). doi:10.1001/jama.279.15.1200
2. Edwards, I.R., Aronson, J.K.: Adverse drug reactions: definitions, diagnosis, and management. Lancet, **356**(9237), 1255–1259 (2000). doi:10.1016/S0140-6736(00)02799-9
3. Chai, G., Governale, L., McMahon, A.W., Trinidad, J.P., Staffa, J., Murphy, D.: Trends of outpatient prescription drug utilization in US children, 2002–2010. Pediatrics, **130**(1), 23–31 (2012). doi:10.1542/peds.2011-2879
4. Lu, Z.: Information technology in pharmacovigilance: benefits, challenges, and future directions from industry perspectives. Drug, Healthcare Patient Safety, 35–45 (2009). https://doi.org/10.2147/dhps.s7180
5. Giangreco, N.P., Tatonetti, N.P.: A database of pediatric drug effects to evaluate ontogenic mechanisms from child growth and development. Med, **3**(8) 579–595 (2022). doi:10.1016/j.medj.2022.06.001
6. Zorych, I., Madigan, D., Ryan, P., Bate, A.: Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat. Methods Med. Res. **22**(1), 39–56 (2013). doi:10.1177/0962280211403602
7. Williams, K., et al.: Standard 6: age groups for pediatric trials. Pediatrics, **129**(3), S153–S160 (2012). doi:10.1542/peds.2012-0055I
8. DuMouchel, W., Harpaz, R.: Regression-Adjusted GPS Algorithm (RGPS). Oracle (2012)
9. DuMouchel, W.: Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. American Stat. **53**(3), 177–190 (1999). doi:10.2307/2686093

10. DuMouchel, W., et al.: Antipsychotics, glycemic disorders, and life-threatening diabetic events: a Bayesian data-mining analysis of the FDA adverse event reporting system (1968–2004). Ann. Clin. Psychiatry, **20**(1) 21–31 (2008). doi:10.1080/10401230701844612
11. Osokogu, O.U., et al.: Pediatric drug safety signal detection: a new drug–event reference set for performance testing of data-mining methods and systems. Drug Safety, **38**(2) 207–217 (2015). doi:10.1007/s40264-015-0265-0

Country Prevalence and Problematic Gambling Among European Adolescents

Elisa Benedetti[1], Gabriele Lombardi[2(✉)], Rodolfo Cotichini[1], Sonia Cerrai[1], Marco Scalese[1], and Sabrina Molinaro[1]

[1] National Research Council, Institute of Clinical Physiology (CNR-IFC), via Moruzzi, 1, 56124 Pisa (PI), Italy
[2] Department of Statistics, Informatics and Applications "G. Parenti", University of Florence, viale Morgagni, 59, 50134 Firenze (FI), Italy
gabriele.lombardi@unifi.it

Abstract. This study investigates the likelihood of engagement in gambling among European adolescents and examines the prevalence of gambling involvement and problem gambling in various countries. The research utilizes data from the European School Survey Project on Alcohol and Other Drugs (ESPAD), employing nationally representative samples of adolescents from 33 European countries. The study reveals regional patterns, with Southern European and Balkan countries exhibiting high prevalence, and Scandinavian and Central European countries displaying it low. The Heckprobit model identifies three distinct profiles of gamblers within each subgroup of countries. The first stage estimates the likelihood of having engaged in gambling, while the second addresses the likelihood of becoming at risk of problematic gaming. Environmental and contextual factors play a more significant role in initiating gambling behavior, while individual characteristics have a greater impact on the likelihood of being at risk. Among various types of games, slot machines emerge as the most significant trigger for gambling behaviors.

Keywords: Gambling · European countries · Adolescents · Heckprobit

1 Introduction

Youth gambling has become an established public health issue due to the high prevalence of gambling involvement and problematic gambling reported in many countries. Problem gambling is a general term that incorporates sub-clinical conditions describing gambling behaviour that results in adverse consequences for individuals, families, and communities and that develops through the interaction of several social, psychological and biological factors. Previous research suggests a prevalence of past-year problem gambling ranging between 0.12% and 5.8% in the general population [1]. The estimated prevalence among youth worldwide is much higher, reaching approximately 0.2–12.3% of them who meet diagnostic criteria for problem gambling [2]. Despite underage gambling is not legal in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 189–195, 2025.
https://doi.org/10.1007/978-3-031-64431-3_32

European countries, the continued expansion of gambling characterised by the legalisation and liberalisation of national gambling markets [3] raises concerns regarding the increased risks for young people stemming from the high availability of multiple gambling products and online gambling. However, in many countries empirical research into adolescent gambling and problem gambling is still limited. Furthermore, the differences among assessment instruments to measure its prevalence used so far have hampered a full overview of the European continent [2]. Cross-sectional studies of adolescents provided the evidence base for the identification of early risk factors, defined in this context as conditions associated to an increased likelihood of problem gambling. Individual-level risk factors include young age, male gender, lower socio-economic status, school difficulties and poor academic performance, psychological distress, impulsivity, sensation seeking and number of gambling activities in the last 12 months. Early relation- and community-level risk factors are peer gambling, parental problem gambling, inconsistent parental discipline, family problems, parents providing money to gamble or buying tickets for games [4], as well as gambling availability, gambling marketing and cultural gambling norms [5]. Despite the proven association of these factors with the development of problem gambling behaviour, many adolescents exposed to them do not develop gambling problems. This suggests that they might influence the likelihood of gambling engagement and the development of a problematic behaviour to a different extent, as well as that there are a number of other factors at work playing a compensatory or protective role. These include interpersonal skills, social competence, support and bonding, school connectedness, parental monitoring and family cohesion [5]. This research is aimed at investigating the different influence that individual-, relation- and environmental-level factors have in triggering the engagement in gambling and the development of problem gambling behaviour among adolescents. In doing so, it investigates geographical differences in adolescent gambling practice and how they reflect in a different role played by the investigated factors. To our knowledge, this is the first study investigating these issue on comparable data from nationally representative samples of adolescents in 33 European countries.

2 Matherial and Methods

2.1 Data

Data are drawn from the European School Survey Project on Alcohol and Other Drugs (ESPAD), which every four years - since 1995 - administeres standardized, in-school, self-reported, cross-sectional surveys on substance use to respondents turning age 16 in the survey year. The sample (n = 85,420) comes from 33 countries that participated in the 2019 data collection and included relevant gambling questions. The study methodology used nationally representative samples of randomly selected classes/schools for the cohort of students turning age 16 in the survey year. The questionnaire collects information on demographics, family environment, gambling and engagement in several risky behaviours. The

proportion of selected classes/schools participating in the survey was 85% on average, while 86% was the mean percentage of students who responded [6].

2.2 Measures

To determine the dependent variable of gambling, students were asked about the frequency of their gambling activity in general and the types of games they played (slot machines, cards or dice, lotteries, or betting on sports/animals) in the past 12 months. Gamblers were defined as those who played for money on at least one of the four included games. The dependent variable of problem gambling was based on the Consumption Screen for Problem Gambling (CSPG) developed by Rockloff [7], which measures gambling frequency, time spent on gambling, and gambling intensity. Those who scored 4 or more points on the CSPG were considered as being at risk of problem gambling [7].

The analysis included several covariates: gender, perceived family and friends' support, missed school days, highest parental education, self-reported family well-off status compared to other families, parental monitoring indicator, and information on how often parents give money to their children. Furthermore, to estimate the popularity of gambling among adolescents within countries, we created the Individual Gambling Product Index (I-GPI). This index measures the quantity and frequency of gambling for each type of game among the peers of our subjects in their respective countries, thus avoiding reflection bias [8]:

$$\text{I-GPI}_{i,g} = \begin{cases} \frac{N_{c,g,ans2}+24\times N_{c,g,ans3}+104\times N_{c,g,ans4}}{(N-1)_c} & \text{if } ans1_i = \text{Yes}, \\ \frac{(N-1)_{c,g,ans2}+24\times N_{c,g,ans3}+104\times N_{c,g,ans4}}{(N-1)_c} & \text{if } ans2_i = \text{Yes}, \\ \frac{N_{c,g,ans2}+24\times (N-1)_{c,g,ans3}+104\times N_{c,g,ans4}}{(N-1)_c} & \text{if } ans3_i = \text{Yes}, \\ \frac{N_{c,g,ans2}+24\times N_{c,g,ans3}+104\times (N-1)_{c,g,ans4}}{(N-1)_c} & \text{if } ans4_i = \text{Yes}, \end{cases} \quad (1)$$

where, c represents the country, g the type of game, ans_1 indicates if certain subjects have never played a certain game, ans_2 indicates if they played it only once in the last 12 months (i.e. 1 day), ans_3 indicates if they played it at least twice per month (i.e. 12 d), ans_4 indicates if they played it at least twice per week (i.e. 104 d). N represents the number of individuals in our sample. The Individual Gambling Product Index (I-GPI) should be interpreted as an indicator of the average frequency of gambling on a particular game in a specific country, excluding the specific subject the indicator refers to. In other words, the I-GPI provides an estimate of the popularity of each game among adolescents within countries, based on the reported frequency of gambling by their peers.

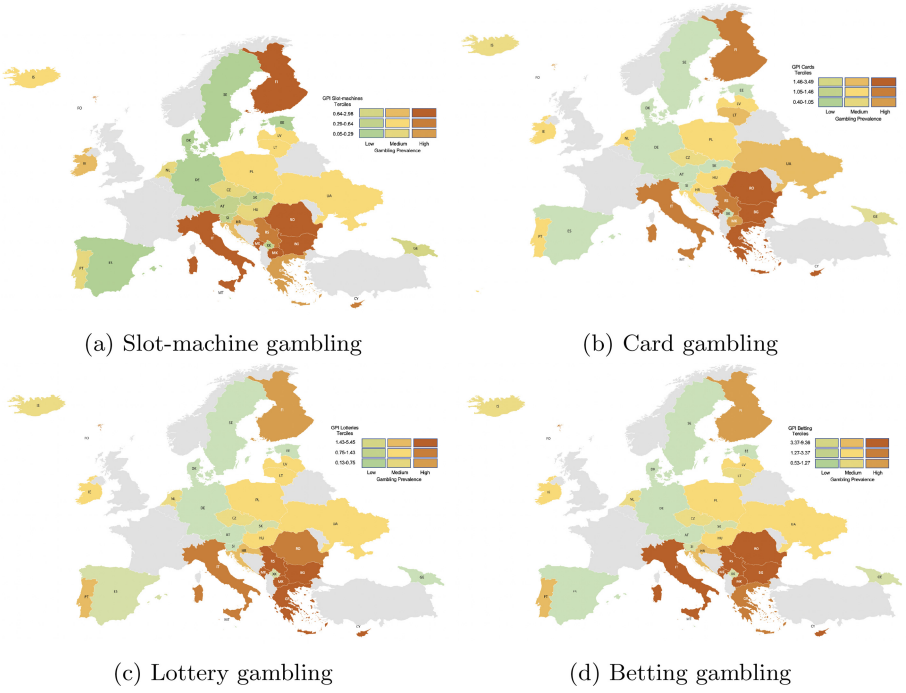
2.3 Empirical Strategy

The analysis is conducted through a probit model with sample selection correction [9]. In particular, in the selection equation all the individual variables are included in order to determine how they influence the probability of being or not

a gambler. Since we are analysing adolescents, it is difficult to strictly determine if they are not-players or they have just not played yet. Therefore, if we want to determine the probability of becoming a problematic gambler we should not consider those who never played any kind of game. As in the standard Heckman model, the necessity emerges of identifying some instrumental variables, which should be relevant for the selection equation, but not for the main equation. For this purpose, we employ the four standardized I-GPIs as environmental variables.

3 Results and Discussion

Figure 1 depicts an overview of Europe based on the prevalence level of gambling of each country and gambling patterns for each game of chance. This has been done by grouping the 33 ESPAD countries into three gambling prevalence sets (high, medium and low) based on tertiles of the N-GPIs. Analysing the four maps, a regional pattern can be noticed where in the high-prevalence group we find mainly Southern European and Balkan countries, in the medium-prevalence group mainly Eastern ones, and in the low-prevalence group predominantly Scandinavian and Central European countries.



(a) Slot-machine gambling

(b) Card gambling

(c) Lottery gambling

(d) Betting gambling

Fig. 1. Gambling prevalence groups and N-GPIs

Table 1. Estimation results by country prevalence groups

	Low Prevalence		Medium Prevalence		High Prevalence	
	NG vs G	Not A-R/AR	NG vs G	Not A-R/AR	NG vs G	Not A-R/AR
	(4.1)	(4.2)	(5.1)	(5.2)	(6.1)	(6.2)
Female	−0.439 * **	−0.812 * **	−0.470 * **	−0.325	−0.646 * **	0.399***
	(0.057)	(0.073)	(0.057)	(0.676)	(0.029)	(0.032)
FamSI	−0.021 * **	−0.045 * **	−0.028*	−0.022	−0.035 * **	0.027***
	(0.005)	(0.018)	(0.007)	(0.015)	(0.008)	(0.010)
FriSI	0.003	0.004	0.018***	−0.015	0.037***	−0.032 * **
	(0.005)	(0.018)	(0.007)	(0.015)	(0.008)	(0.010)
School Missed (Ref: 0 d)						
1–2 days	0.102***	0.074	0.177***	−0.135	0.184***	−0.139 * **
	(0.009)	(0.062)	(0.024)	(0.123)	(0.027)	(0.030)
3–5 days	0.179***	0.210***	0.242***	−0.002	0.337***	−0.269 * **
	(0.028)	(0.066)	(0.022)	(0.224)	(0.026)	(0.027)
5 d or more	0.291***	0.382***	0.318***	0.111	0.430***	−0.285 * **
	(0.044)	(0.068)	(0.025)	(0.338)	(0.040)	(0.026)
Highest Parental Education (Ref: Less than High School (HS))						
HS	0.048	0.049	0.050	−0.044	0.155***	−0.137 * **
	(0.035)	(0.056)	(0.032)	(0.041)	(0.035)	(0.039)
More than HS	−0.014	0.028	0.010	−0.089	0.147***	−0.147 * **
	(0.037)	(0.057)	(0.038)	(0.088)	(0.033)	(0.029)
Family Well off (Ref: Less off)						
About the same	−0.024	−0.004	−0.098 * **	−0.057	0.048	−0.060 * *
	(0.041)	(0.083)	(0.027)	(0.125)	(0.033)	(0.030)
Better off	0.066	0.116	0.031	−0.017	0.085*	−0.066
	(0.043)	(0.089)	(0.033)	(0.052)	(0.047)	(0.044)
Parental Monitoring (Ref: About Always)						
Sometimes	0.237***	0.114**	0.202***	−0.081	0.244***	−0.186 * **
	(0.028)	(0.052)	(0.013)	(0.134)	(0.036)	(0.027)
About Never	0.335***	0.430***	0.348***	0.015	0.351***	−0.157 * **
	(0.041)	(0.065)	(0.036)	(0.317)	(0.063)	(0.047)
Parents give money (Ref: Seldom/Never)						
Often/Sometimes	0.050**	−0.026	0.079***	−0.085 * **	0.094***	−0.107 * **
	(0.021)	(0.044)	(0.026)	(0.022)	(0.029)	(0.023)
Almost Always	0.112***	0.097**	0.174***	0.073	0.125***	−0.085 * *
	(0.028)	(0.044)	(0.037)	(0.151)	(0.037)	(0.035)
I-GPI$_{slot}$	0.311	–	0.056	–	0.031**	–
	(0.242)	–	(0.120)	–	(0.015)	–
I-GPI$_{cards}$	−0.053	–	0.078	–	0.027	–
	(0.098)	–	(0.073)	–	(0.021)	–
I-GPI$_{lotteries}$	−0.167	–	−0.117	–	0.017	–
	(0.126)	–	(0.094)	–	(0.011)	–
I-GPI$_{betting}$	0.010	–	−0.048	–	−0.005	–
	(0.079)	–	(0.066)	–	(0.007)	–
Constant	−0.996 * **	−1.954 * **	−0.950 * **	0.369	−0.936 * **	0.992***
	(0.155)	(0.121)	(0.078)	(1.269)	(0.073)	(0.058)
No. Obs.	26,626		28,377		30,546	
Log-Likelihood	−12106.41		−14395.92		−20636.58	
Err. Corr.	0.971***	(0.016)	−0.679	(0.634)	−0.993 * **	(0.003)

Note: the table reports coefficient estimates and standard errors in parentheses.
Standard errors are clustered at country level.
Survey data are weighted. =”* p<0.1 ** p<0.05 *** p<0.01”

Table 1 summarizes the results of three different Heckprobit models based on the three subsample of countries (i.e. Low, Medium and High Prevalence of gaming). For each of the three regression the first column identifies the sample selection step between gamblers and non-gamblers (NG vs G) and the second column the main regression for the likelihood of being at-risk of problematic gambling behavior (Not A-R/AR).

This exploratory analysis provided an insight into the individual and environmental factors driving the engagement in gambling and the development of a problematic gambling behaviour among adolescents in Europe. We asked ourselves whether the popularity of gambling in a country constitutes a driver in itself for the development of at-risk gambling behavior and if there are specific games acting tiggering more than others. We observed a high variability among countries in terms of the prevalence of gambling engagement and of problem-gambling. The preliminary results portray distinct profiles of both adolescent gamblers and problem gamblers, highlighting the significant role played by the proliferation of slot machines in triggering gambling engagement. Environmental and contextual factors seem to be much more important in determining the chances of starting to gamble, rather than those of developing a problematic behavior. In the latter case, individual characteristics of subjects and of their own immediate setting composed by family and friends seem to matter more, suggesting that both unconditioned trust and excessive control, summarized by the frequency of both parental monitoring and provision of money, can be factors increasing the chances of progressing into a problematic gambling behaviour. Regarding gender differences, we also found evidence compatible with the so-called telescoping effect hypothesis [10]: women are less likely to start gambling but progress to problem gambling faster. Indeed, females exhibit a lower risk of problematic behavior only in Low Prevalence countries, while they face a higher risk in High Prevalence ones, and no gender effect emerges in the Medium Prevalence country group.

References

1. Calado, F., Griffiths, M.D.: Problem gambling worldwide: an update and systematic review of empirical research (2000–2015). J. Behav. Addict. **5**, 592–613 (2016)
2. Calado, F., Griffiths, M.D.: Prevalence of adolescent problem gambling: a systematic review of recent research. J. Gambl. Stud. **33**, 397–424 (2017)
3. Aimo, N., Bassoli, M., Marionneau, V.: A scoping review of gambling policy research in Europe. Int. J. Soc. Welfare 1–16 (2023)
4. Reith, G., Dobbie, F.: Beginning gambling: the role of social networks and environment. Addict. Res. Theor. **19**, 1483–493 (2011)
5. Dowling, N.A., Merkouris, S.S., Greenwood, C.J., Oldenhof, E., Toumbourou, J.W., Youssef, G.J.: Early risk and protective factors for problem gambling: a systematic review and meta-analysis of longitudinal studies. Clin. Psychol. Rev. **51**, 109–124 (2017)
6. ESPAD Group (2020).: ESPAD Report 2019: Results from the European School Survey Project on Alcohol and Other Drugs. Publications Office of the European Union, Luxembourg. ISBN: 978-92-9497-546-1

7. Rockloff, M.: Validation of the consumption screen for problem gambling (CSPG). J. Gambl. Stud. **28**, 207–16 (2012)
8. Manski, C.F.: Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. **60**(3), 531–542 (1993)
9. Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**(1), 153–161 (1979)
10. Gartner, C., Bickl, A., Härtl, S., Loy, J. K., Häffner, L.: Differences in problem and pathological gambling: a narrative review considering sex and gender. J. Behav. Addictions (2022)

Stochastic Variational Inference for Structured Bayesian Hierarchical Models

Massimo Bilancia[1(✉)] and Andrea Nigri[2]

[1] Department of Precision and Regenerative Medicine and Jonian Area (DiMePRe-J), University of Bari Aldo Moro, Bari, Italy
massimo.bilancia@uniba.it
[2] Department of Economics, Mangement and Territory (DEMeT), University of Foggia, Foggia, Italy

Abstract. In this paper we consider stochastic variational inference for finite mixtures of Dirichlet-Multinomial distributions. By exploiting simple hypotheses concerning the full conditional distributions of the hierarchical model and the distributions of the variational parameters, a gradient ascent algorithm can be derived that under the Robbins-Monro conditions converges to a local maximum of the surface approximating the posterior distribution.

Keywords: Dirichlet-Multinomial finite mixture model · stochastic variational inference

1 Introduction

The class of Bayesian hierarchical models we discuss in this paper is of the form:

$$p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\delta) = p(\boldsymbol{\omega}|\delta) \prod_{i=1}^{n} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\omega}), \tag{1}$$

where $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ and $\boldsymbol{z}_{1:n}$ is a set of latent variables defined accordingly [3]. The conditional dependencies made explicit in (1) illustrate the difference between local latent variables (one for each observed data point) and global latent variables (true parameters). In particular, the data and the local latent variables are conditionally independent given the global parameters. The specific example of (1) that we will refer to is the following Dirichlet-Multinomial finite mixture model for positive count data:

$$\boldsymbol{y}_i|\mathbb{B}, \boldsymbol{z}_i \overset{\text{ind.}}{\sim} \mathsf{Multinomial}_p(\boldsymbol{\beta}_j), \quad i = 1, 2, \ldots, n, \tag{2}$$

$$\boldsymbol{z}_i|\boldsymbol{\lambda} \overset{\text{ind.}}{\sim} \mathsf{Multinoulli}_k(\boldsymbol{\lambda}), \quad i = 1, 2, \ldots, n, \tag{3}$$

$$\boldsymbol{\beta}_j|\theta \overset{\text{ind.}}{\sim} \mathsf{Dirichlet}_p(\mathbb{1}_p\theta), \quad j = 1, 2, \ldots, k, \tag{4}$$

$$\boldsymbol{\lambda}|\alpha \sim \mathsf{Dirichlet}_k(\mathbb{1}_k\alpha). \tag{5}$$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 196–201, 2025.
https://doi.org/10.1007/978-3-031-64431-3_33

2 Fixed Form Variational Inference

In the following, we will refer to the formulation (1) for reasons of maximum generality. Variational inference approximates the intractable posterior with the mean-field family:

$$q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu}) = q(\boldsymbol{\omega}|\boldsymbol{\zeta}) \prod_{i=1}^{n} q(\boldsymbol{z}_i|\boldsymbol{\gamma}_i), \tag{6}$$

with $\boldsymbol{\nu} = (\boldsymbol{\omega}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$. The decoupling between the global parameters and the local variables makes the approximate posterior inference tractable. The quality of the approximation is determined by minimizing the reverse Kullback-Leibler (KL) divergence between the variational distribution and the true posterior distribution:

$$\underset{q}{\operatorname{argmin}}\, \mathsf{KL}\left(q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu}) || p(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{y}_{1:n}, \delta)\right). \tag{7}$$

Defining the evidence lower-bound (ELBO) as follows:

$$\mathsf{ELBO}(q) = \mathsf{E}_q\left[\log p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\delta)\right] - \mathsf{E}_q\left[\log q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu})\right], \tag{8}$$

as a consequence of a standard proof using Jensen's inequality and the concavity of the logarithm function, the fundamental identity of variational inference follows [3]:

$$\log p(\boldsymbol{y}_{1:n}|\delta) = \mathsf{ELBO}(q) + \mathsf{KL}\left(q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu}) || p(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{y}_{1:n}, \delta)\right). \tag{9}$$

3 The Natural Gradient

Maximizing the ELBO with respect to the variational parameters $\boldsymbol{\nu}$ on a subset of Euclidean space may not be the most appropriate strategy to ensure fast convergence to a local maximum point. The maximization of $\mathcal{L}_\nu \equiv \mathsf{ELBO}$ is a functional maximization problem on the statistical manifold parameterized by the variational distribution $q_{\boldsymbol{\nu}} \equiv q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega})$, where it is well known from the methods of information geometry that $\mathsf{KL}(q_{\boldsymbol{\nu}}||q_{\boldsymbol{\nu}'})$ is a suitable Riemannian metric over this manifold that conveys the information about the local geometry in a neighborhood of a fixed member of the family $q_{\boldsymbol{\nu}}$. With a simple demonstration that does not use any notions of Riemannian geometry, it can be shown that the direction of maximum ascent, i.e. the natural gradient vector on the statistical manifold, has the following approximated expression in a small neighborhood of $q_{\boldsymbol{\nu}}$ (up to second order):

$$\nabla_{\boldsymbol{\nu}} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} \approx \mathbb{F}_{\boldsymbol{\nu}}^{-1} \nabla_{\boldsymbol{\nu}} \mathcal{L}_{\boldsymbol{\nu}}, \tag{10}$$

where $\mathbb{F}_{\boldsymbol{\nu}}$ is the expected Fisher information matrix of the parametric family $q_{\boldsymbol{\nu}}$:

$$\mathbb{F}_{\boldsymbol{\nu}} = \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\left(\nabla_{\boldsymbol{\nu}} q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega})\right)\left(\nabla_{\boldsymbol{\nu}} q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega})\right)^{\top}\right]. \tag{11}$$

4 Stochastic Variational Inference

With fixed form variational inference, the coordinate ascent update equation for the local variational parameters are (for $t = 0, 1, 2, \ldots$) [1]:

$$\boldsymbol{\gamma}_i^{(t+1)} = \mathsf{E}_{q_{\boldsymbol{\zeta}^{(t)}}} \left[\eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega}) \right], \quad i = 1, 2, \ldots, n. \tag{12}$$

which are coupled with the update equation of the global variation parameter:

$$\boldsymbol{\zeta}^{(t+1)} = \mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \delta) \right]. \tag{13}$$

The equations for updating the local variational parameters must be evaluated through the function $\eta_\ell(\cdot)$ for the entire data set at hand.

By reformulating the update equation (13) of the global variational parameter as a natural gradient ascent with step size $\rho^{(t)}$, we obtain:

$$\begin{aligned} \boldsymbol{\zeta}^{(t+1)} &= \boldsymbol{\zeta}^{(t)} + \rho^{(t)} \left(\mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \delta) \right] - \boldsymbol{\zeta}^{(t)} \right) = \\ &= (1 - \rho^{(t)}) \boldsymbol{\zeta}^{(t)} + \rho^{(t)} \mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \delta) \right] = \\ &= (1 - \rho^{(t)}) \boldsymbol{\zeta}^{(t)} + \rho^{(t)} \widehat{\boldsymbol{\zeta}}^{(t)}, \end{aligned} \tag{14}$$

where:

$$\widehat{\boldsymbol{\zeta}}^{(t)} = \mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \delta) \right], \tag{15}$$

With variational inference [4], the idea to improve scalability is to replace this expected value by a noisy but unbiased and cheap to compute estimate based on a single sampled data point $\boldsymbol{z}_s$, where:

$$s \sim \mathsf{Uniform}\,(1, 2, \ldots, n),$$

and:

$$\boldsymbol{z}_s^n = (\overbrace{\boldsymbol{z}_s, \boldsymbol{z}_s, \ldots, \boldsymbol{z}_s}^{n \text{ times}}).$$

The natural gradient estimate based on this surrogate dataset is:

$$\widehat{\nabla}_{\boldsymbol{\zeta}}^{s,n} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} = \mathsf{E}_{q_{\boldsymbol{\gamma}}} \left[\eta_g(\boldsymbol{y}_s^n, \boldsymbol{z}_s^n, \delta) \right] - \boldsymbol{\zeta}, \tag{16}$$

and thus the algorithm runs through an optimized local parameter using a single sampled data point:

$$\boldsymbol{\gamma}_s^{(t+1)} = \mathsf{E}_{q_{\boldsymbol{\zeta}^{(t)}}} \left[\eta_\ell(\boldsymbol{y}_s, \boldsymbol{\omega}) \right], \tag{17}$$

and a natural gradient ascent step based on its noisy estimate:

$$\widehat{\boldsymbol{\zeta}}^{(t)} = \mathsf{E}_{q_{\boldsymbol{\gamma}}^{(t+1)}} \left[\eta_g(\boldsymbol{y}_s^n, \boldsymbol{z}_s^n, \delta) \right]. \tag{18}$$

which in turn must be inserted in (14).

Convergence to a local optimum is guaranteed as long as the Robbins-Monro conditions for the step size are satisfied:

$$\sum_{t=1}^{+\infty} \rho^{(t)} = \infty, \quad \sum_{t=1}^{+\infty} \left(\rho^{(t)} \right)^2 < \infty. \tag{19}$$

5 Stochastic Variational Inference for the Dirichlet-Multinomial Finite Mixture Model

In [4], a general framework framework for variational inference in exponential families is introduced, referred to as conditional conjugacy, which is based on the existence of a conjugacy relationship in (1) between the prior distribution of the global variable $\boldsymbol{\omega}$ and the joint distribution $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ relative to the local context of the ith observation, where both distributions belong to the exponential family. These assumptions are necessary to clarify the actual shape of the global natural parameter $\eta_g \equiv \eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \delta)$ in a general setting.

However, based on a more restricted and easier to handle set of assumptions, we are able to make explicit the expected values that appear in the update equations (14) and (17). In particular, it is necessary that:

- The full conditional of the global parameter $\boldsymbol{\omega}$ is the exponential family.
- The full conditional of each local hidden variable $\boldsymbol{z}_i$ is a canonical exponential family.
- The global component $\boldsymbol{\omega}$ of the mean field family (6) is in the same exponential family as its full conditional.
- The local component $\boldsymbol{z}_i$ is in the same exponential family as its full conditional.

Under these assumptions, it can be proven that stochastic variational inference for the Dirichlet-Multinomial finite mixture model (2)–(5) evolves as shown in the Algorithm 1 (details can be found in [2]).

6 Example

We applied stochastic variational inference to the binary text categorization of a document corpus to compare it with the coordinate ascent variational inference (CAVI). The dataset used is a subset from the collection Reuters 21578. The corpus used for the experiment consists of $n = 70$ documents, 50 of which belong to the category `acq` (containing news about corporate takeovers) and 20 to the category `crude` (containing news about crude oil), with a clear imbalance between the two classes. To convert the text into a BOW, the raw data was preprocessed with R 4.2.3 using standard steps for natural language processing.

In both cases (CAVI and stochastic variational inference as described in the Algorithm 1, from now on SVI) we have used $k = 2$ and the default values $\alpha = 1$ and $\theta = 5/k$. This setting is neutral with respect to α and weakly informative with respect to θ. Although variational inference explores a single mode of the posterior surface, each run of the algorithm converges to a different local maximum of ELBO and multiple runs are required to find the optimal one. Of course, this procedure is not unproblematic from a computational point of view and the solution obtained may be suboptimal, since the local maxima examined do not necessarily contain the global maximum. Examples of ELBO trajectories towards a local maximum for multiple runs are shown in Fig. 1.

Algorithm 1. Stochastic variational inference for the Dirichlet-Multinomia finite mixture model (2)–(5)

Input: Data $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$; number of components k; prior hyperparameters $\theta, \alpha > 0$; variational families $q_{\boldsymbol{\gamma}_i}(\boldsymbol{z}_i)$, $q_{\boldsymbol{\phi}_j}(\boldsymbol{\beta}_j)$ for $j = 1, 2, \ldots, k$, $q_{\boldsymbol{\eta}}(\boldsymbol{\lambda})$; step size $\rho^{(t)}$ for $t = 1, 2, \ldots$

Initialize: Variational parameters γ_{ij} for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$; $\phi_{j\ell}$ for $j = 1, 2, \ldots, k$, $\ell = 1, 2, \ldots, p$; η_j for $j = 1, 2, \ldots, k$ (initialize randomly in $t = 0$).

Output: Optimized variational densities.

while the ELBO has not converged, for $t = 1, 2, \ldots$ **do**:
  Sample $s \sim \mathsf{Uniform}(1, 2, \ldots, n)$
  Update local variational parameters:
  for $j = 1, 2 \ldots, k$ **do**
    $\gamma_{sj} = \exp\left\{\sum_{\ell=1}^{p} y_{s\ell} \mathsf{E}_q\left[\log \beta_{j\ell}\right]\right\} \exp\left\{\mathsf{E}_q\left[\log \lambda_j\right]\right\}$
    $\gamma_{sj} \leftarrow \frac{\gamma_{sj}}{\sum_{h=1}^{k} \gamma_{sh}}$
  end for
  Intermediate global variational parameters:
  for $j = 1, 2, \ldots, k$ **do**
    for $\ell = 1, 2, \ldots, p$ **do**
      $\widehat{\phi}_{j\ell} = \theta + n y_{s\ell} \gamma_{sj}$
    end for
  end for
  for $j = 1, 2, \ldots, k$ **do**
    $\widehat{\eta}_j = \alpha + n\gamma_{sj}$
  end for
  Update global variational parameters:
  for $j = 1, 2, \ldots, k$ **do**
    for $\ell = 1, 2, \ldots, p$ **do**
      $\phi_{j\ell} \leftarrow \left(1 - \rho^{(t)}\right) \phi_{j\ell} + \rho^{(t)} \widehat{\phi}_{j\ell}$
    end for
  end for
  for $j = 1, 2, \ldots, k$ **do**
    $\eta_j \leftarrow \left(1 - \rho^{(t)}\right) \eta_j + \rho^{(t)} \widehat{\eta}_j$
  end for
end while

The results obtained with SVI are reassuring both in terms of scalability and the quality of the estimated solutions (Table 1). To compare the two algorithms, we set by visual inspection an appropriate number of iterations, denoted as `maxiter`, which guarantees convergence for all considered runs. For CAVI we set `maxiter` = 50, while for SVI `maxiter` = 350. The comparison was not only based on the computation times, but also on the evaluation of the quality of the unsupervised posterior classification obtained via the classical allocation rule that applies to a 0/1 loss function, where $\boldsymbol{\gamma}_i^{\star}$ denotes the optimized variational hyperparameter:

$$j(i)^{\mathsf{MAP}} = \underset{j}{\operatorname{argmax}}\, q_{\boldsymbol{\gamma}_i^{\star}}(z_{ij} = 1), \quad i = 1, 2, \ldots, n. \tag{20}$$

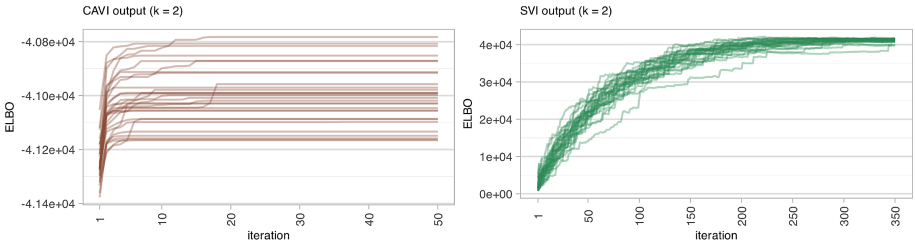


Fig. 1. 30 randomly selected trajectories of the CAVI algorithm applied to a subset of Reuters 21578 data with $k = 2$, `maxiter` $= 100$ and `nruns` $= 50$. Right: As in the left panel, except that the trajectories were obtained using SVI with `maxiter` $= 350$.

Table 1. Total, per run and per iteration computation times, unsupervised classification accuracy (Acc) obtained with the allocation rule (20), and the adjusted Rand index (ARI) for the two algorithms (CAVI and SVI), corresponding to multiple choices of `nruns`. The results were obtained on a subset of the collection Reuters 21578.

Inference	`maxiter`	`nruns`	t (total)	t (avg. per run)	t (avg. per iter.)	Acc	ARI
CAVI							
	50	10	10.77 s	1.077 s	0.022 s	67.14	0.48
	50	20	21.29 s	1.065 s	0.021 s	84.29	44.36
	50	50	53.56 s	1.071 s	0.021 s	81.43	38.22
	50	100	108.47 s	1.085 s	0.021 s	92.86	72.59
SVI							
	350	10	33.05 s	3.305 s	0.009 s	94.29	77.26
	350	20	66.19 s	3.310 s	0.009 s	92.86	72.14
	350	50	165.23 s	3.305 s	0.009 s	97.14	88.39
	350	100	331.71 s	3.317 s	0.009 s	94.29	77.01

References

1. Bilancia, M., Di Nanni, M., Manca, F., Pio, G.: Variational bayes estimation of hierarchical dirichlet-multinomial mixtures for text clustering. Comput. Statistics 1–37 (2023). https://doi.org/10.1007/s00180-023-01350-8
2. Bilancia, M., Nigri, A.: Efficient stochastic variational inference for the Dirichlet-Multinomial finite mixture model, submitted (2024)
3. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**, 859–877 (2017). https://doi.org/10.1080/01621459.2017.1285773
4. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. J. Mach. Learn. Res. **14**, 1303–1347 (2013)



Advances in Intermittent Demand Forecasting

Parvaneh Rafieisangari and Luisa Bisaglia(✉)

University of Padova, Via Cesare Battisti 241, 35121 Padua, PD, Italy
parvaneh.rafieisangari@phd.unipd.it, bisaglia@stat.unipd.it

Abstract. Time series of intermittent demand display an erratic pattern that differs from other demand types, with a higher occurrence of zero demands. For this reason, forecasting intermittent demand proves challenging. Most existing methods for forecasting time series of intermittent demand with high intermittency data are unsuitable for generating accurate forecasts. Moreover, we have to take into account the integer nature of data. In particular, given the discrete nature of the data, in this work we suggest employing integer-valued autoregressive models for forecasting. Through empirical evaluation on real dataset, we illustrate significant statistical improvements in the generated forecasts.

Keywords: Intermittent Demand Data · Forecasting · Integer-valued autoregressive processes · Zero-Inflated distribution

1 Introduction

According to Silver et al. [1], intermittent demand is characterized by its infrequency, where the average time between transactions is significantly greater than the unit period, the latter representing the interval for updating forecasts. Johnston and Boylan's [2] suggestion, on the other hand, categorizes demand as intermittent when the average time between successive demands exceeds 1.25 review intervals. Croston [3], raised concerns about the suitability of traditional forecasting techniques like Simple Moving Averages (SMA) and Single Exponential Smoothing (SES) for items with slow-moving demand patterns. He quantitatively illustrated the bias associated with SES when updating forecasts following a demand occurrence. To address these issues, he proposed a novel method that utilizes separate Single Exponential Smoothing (SES) forecasts for demand size and the time interval between two consecutive demands, both employing the same smoothing constant α. This method was designed specifically to better handle intermittent demand situations. Many subsequent approaches are essentially variations of the original Croston's method, reflecting its enduring impact. Approaches such as Croston's method may enhance forecast precision in time series characterized by low intermittency. However, its effectiveness diminishes in forecasting highly intermittent time series. Considering the integer nature of the data, we propose to modify Croston's method using the INteger-valued

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 202–207, 2025.
https://doi.org/10.1007/978-3-031-64431-3_34

AutoRegressive (INAR) models for forecasting intermittent demand. The outcomes suggest that using this model it is possible to obtain stronger alignment with the data and better performances in the forecasting process respect to other methods. The remaining sections of the paper are structured as following: Sect. 2 introduces Croston's methods and their variations for forecasting intermittent demand time series. Section 3 delves into forecasting using the integer-valued autoregressive model and its specific instances. Section 4 illustrates the analyses of real data and the corresponding results. Finally, in the Sect. 5, concluding remarks are presented.

2 Croston's Method

Croston's method, introduced by Croston [3], stands as a widely utilized predictive technique specifically tailored for intermittent demand series. Intermittent demand patterns, characterized by sporadic and unforeseeable instances of non-zero demand, present a challenge for traditional forecasting methods. This approach addresses this challenge by dividing the intermittent demand time series into two different time series: one pertaining to non-zero demand occurrences and the other concerning the periods between occurrences. The rationale underpinning this methodology is to prevent an overestimation of demand sizes and to accurately capture the intermittent nature inherent in demand patterns. The formulation of Croston's method comprises two main equations: one for approximating the rate of demand occurrences and another for approximating the intervals between occurrences. Let Z'_t symbolize the estimate of the demand size at time t, T'_t be the estimate of the demand interval at time t and D_t denote the demand for an item at time t. The two equations of the Croston's method are:

1. if $D_t > 0$, then
$$Z'_t = \alpha Z_t + (1-\alpha)Z'_{t-1},$$
where Z_t is the actual demand size at time t, otherwise $Z'_t = Z'_{t-1}$;
2. if $D_t > 0$, then
$$T'_t = \alpha T_t + (1-\alpha)T'_{t-1},$$
where T_t is the actual demand interval at time t, otherwise $T'_t = T'_{t-1}$.

The smoothing parameter α takes values between 0 and 1 and is assumed to be the same for both equations.

By utilizing the updated demand occurrence rate and the adjusted intervals between occurrences, the forecast for the forthcoming period, D'_t, can be computed as:
$$D'_t = \frac{Z'_t}{T'_t}.$$

2.1 Syntetos-Boylan Approximation (SBA)

The effectiveness of Croston's method lies in its capability to adapt the predictive approach in response to the irregular nature of intermittent demand, leading to more precise estimations and projections. Syntetos and Boylan [4] demonstrated that Croston's Method suffers from a positive bias. To mitigate this issue, Syntetos and Boylan [5] proposed a damping factor $\left(1-\frac{\alpha}{2}\right)$, that is directly multiplied by the demand rate as estimated by Croston's approach, $D'_t = \left(1-\frac{\alpha}{2}\right)\frac{Z'_t}{T'_t}$.

2.2 Shale-Boylan-Johnston (SBJ) Approach

Shale et al., [6] proposed another modification, called Shale-Boylan-Johnston (SBJ), by multiplying the factor $\left(1-\frac{\alpha}{2-\alpha}\right)$ with Croston's estimated demand rate, $D'_t = \left(1-\frac{\alpha}{2-\alpha}\right)\frac{Z'_t}{T'_t}$.

2.3 Teunter-Syntetos-Babai (TSB) Approach

Croston's, SBA, and SBJ methods only update demand magnitudes and intervals during periods of positive demand. To tackle this problem, Teunter and Duncan [7] propose a Teunter-Syntetos-Babai (TSB) approach, wherein demand occurrence probabilities are employed to estimate demand magnitudes instead of using demand intervals. Let p'_t symbolize the estimate of the demand occurrence probability at time t, and $p'_t = \alpha p_t + (1-\alpha)p'_{t-1}$, where p_t is the actual demand occurrence probability at time t, then $D'_t = p'_t Z'_t$.

In general, none of the mentioned methods prove successful in forecasting data with high intermittency.

3 Integer-Valued Autoregressive Model (INAR)

A prevalent method for modeling integer-valued autoregressive (INAR) time series involves the use of thinning operators. The most widely adopted thinning operator is the binomial thinning operator, which was initially presented by Steutel and van Harn [8]. Suppose Y is a non-negative random variable that takes integer values, and let $\alpha \in (0,1)$. The binomial thinning operator is $\alpha \circ Y = \sum_{i=1}^{Y} X_i$, where X_i are independent and identically distributed (iid) Bernoulli random variables, independent of Y, with $Pr(X_i = 1) = 1 - Pr(X_i = 0) = \alpha$.

Using the concept of the binomial thinning operator, McKenzie [9] and Al-Osh and Alzaid [10] independently introduced a first order integer-valued autoregressive (INAR(1)) model in the following manner:

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t, \quad \text{if } t \in \mathbf{Z}$$

where $\epsilon_t,\ t \in \mathbf{Z}$, is a sequence of independent and identically distributed (iid) integer-valued random variables with finite mean and variance, ϵ_t and y_{t-i} being independent for all $i \geq 1$.

3.1 Poisson INAR(1) Model

The most popular INAR model uses Poisson innovations and is called Poisson INAR(1) process (PINAR(1)). It assumes that ϵ_t is a sequence of independent Poisson distributed variables with parameter λ, independent of all Y_t. Note that, if $\epsilon_t \sim Po(\lambda)$ then $Y_t \sim Po(\frac{\lambda}{(1-\alpha)})$ (McKenzie [11]). According to Freeland and McCabe [12] the k-step ahead conditional mean and variance of the process are:

$$E(Y_{t+k}|Y_t) = \alpha^k Y_t + \lambda(\frac{1-\alpha^k}{1-\alpha}), \qquad V(Y_{t+k}|Y_t) = \alpha^k(1-\alpha^k)Y_t + \lambda(\frac{1-\alpha^k}{1-\alpha}).$$

To estimate the unknown parameters of the PINAR(1) model it is possible to use Yule-Walker method:

$$\hat{\alpha}_{YW} = \frac{\sum_{t=2}^{n}(y_t-\bar{y})(y_{t-1}-\bar{y})}{\sum_{t=2}^{n}(y_t-\bar{y})^2}, \qquad \hat{\lambda}_{YW} = \bar{y}(1-\hat{\alpha}_{YW})$$

3.2 Zero-Inflated Poisson INAR(1) Model

Since the intermittent demand series are characterized for their high number of zeros, the zero-inflated Poisson distribution seems to be more able to capture this feature of data. A random variable Y is said to be a zero-inflated Poisson distribution with parameters ρ and λ, denoted as $ZIP(\rho,\lambda)$ if its probability mass function (pmf) is defined as

$$P(Y=y) = \begin{cases} \rho + (1-\rho)e^{-\lambda}, & \text{if } y=0 \\ (1-\rho)\frac{\lambda^y e^{-\lambda}}{y!}, & \text{if } y=1,2,\ldots \end{cases}$$

where $\lambda > 0$ and ρ is the probability of extra zeros. Poisson distribution is a special case of zero-inflated Poisson distribution when ρ is 0. The Zero-Inflated Poisson INAR(1) process (ZIPINAR(1)), see Jazi et.al., [13], for further details, assume that ϵ_t is a sequence of independent Zero-Inflated Poisson distributed variables with mean $\mu_\epsilon = \lambda(1-\rho)$ and variance $\sigma_\epsilon^2 = \lambda(1-\rho)(1-\rho\lambda)$, ($\epsilon_t \sim ZIP(\rho,\lambda)$), independent of all Y_t. According to Piancastelli and Barreto-Souza [14] the k-step conditional mean and variance of the process are:

$$E(Y_{t+k}|Y_t) = \alpha^{k+1}Y_t + \mu_\epsilon(\frac{1-\alpha^{k+1}}{1-\alpha})$$

$$V(Y_{t+k}|Y_t) = \alpha^{k+1}(1-\alpha^{k+1})Y_t + \sum_{h=o}^{k}(\alpha^h(1-\alpha^h)\mu_\epsilon + \alpha^{2h}\sigma_\epsilon^2)$$

Again, to estimate the unknown parameters of the ZIPINAR(1) model it is possible to use Yule-Walker method:

$$\hat{\alpha}_{YW} = \frac{\sum_{t=2}^{n}(y_t-\bar{y})(y_{t-1}-\bar{y})}{\sum_{t=2}^{n}(y_t-\bar{y})^2}, \qquad \hat{\lambda}_{YW} = \bar{y}(1-\hat{\alpha}_{YW}) + (1+\hat{\alpha}_{YW})\left(\frac{s_n}{\bar{y}} - 1\right)$$

$$\hat{\rho}_{YW} = 1 - \frac{\bar{y}(1-\hat{\alpha}_{YW})}{\hat{\lambda}_{YW}}, \qquad s_n = \frac{\sum_{t=2}^{n}(y_t-\bar{y})^2}{n}$$

.

4 Real Data Analysis

We assess the effectiveness of the suggested method by employing a real dataset from the Royal Air Force in the United Kingdom. This dataset contains the monthly demand histories (84 months), of 5000 Stock Keeping Units (SKU), from January 1996 to December 2002 (Syntetos et al. [15]; Teunter and Duncan [7]). On average, for all time series, 89.8% of periods experience zero demand, indicating a high level of intermittency. Additionally, in instances where demand is present, most time series exhibit considerable demand variability.

For the evaluation of the forecast performances, we use mean absolute error (MAE) and root mean square error (RMSE), the test set consists of the last 10 months for each series, and we conduct a rolling forecast method (considering horizon of 5 d). This involves generating a forecast, assessing its performance, and then advancing the forecast origin to include the next data point from the test set in the training set. This process is iteratively repeated until all available test data is utilized (Fig. 1). For each forecasting method and time series, we generate 5 forecast horizons within each window, ensuring a comprehensive measurement of performance.

In the context of Table 1, ZIPINAR(1) emerges as the standout performer, boasting the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These results underscore its exceptional forecasting accuracy, positioning ZIPINAR(1) as the optimal choice among the various methods presented in the tables. Its ability to consistently outperform others highlights its efficacy in delivering precise and reliable forecasts for the given data.



Fig. 1. Rolling forecast method

Table 1. Croston vs. INAR-type models: forecasting performance

	RMSE							MAE						
	W1	**W2**	**W3**	**W4**	**W5**	**W6**	**Mean**	**W1**	**W2**	**W3**	**W4**	**W5**	**W6**	**Mean**
CROSTON	1.48	1.46	1.45	1.44	1.44	1.34	1.44	1.43	1.41	1.39	1.38	1.38	1.34	1.39
SBA	1.51	1.46	1.53	1.44	1.35	1.29	1.43	1.45	1.41	1.48	1.38	1.29	1.29	1.38
SBJ	1.48	1.59	1.53	1.44	1.44	1.33	1.47	1.43	1.53	1.48	1.38	1.38	1.33	1.42
TSB	1.50	1.48	1.46	1.46	1.30	1.42	1.44	1.45	1.43	1.41	1.41	1.24	1.42	1.39
ZIPINAR(1)	0.43	0.42	0.43	0.44	0.41	0.00	0.32	0.20	0.20	0.20	0.22	0.20	0.00	0.14
PINAR(1)	0.44	0.44	0.45	0.44	0.44	0.04	0.37	0.22	0.22	0.20	0.22	0.22	0.04	0.18

5 Conclusion

In this paper, we employed a first-order Integer-valued Autoregressive model with Poisson and Zero-Inflated Poisson innovation for forecasting intermittent demand. The performance of this model was compared with other methods, including Croston and its modifications. Through the application of these methods to a real dataset, our findings indicate that the proposed model exhibits superior forecasting performance.

References

1. Silver, E.A., Pyke, D.F., Peterson, R.: Inventory Management and Production Planning and Scheduling, 3rd edn. John Wiley & Sons, New York (1998)
2. Johnston, F.R., Boylan, J.E.: Forecasting for items with intermittent demand. J. Oper. Res. Soc. **47**(1), 113–121 (1996)
3. Croston, J.D.: Forecasting and stock control for intermittent demands. Oper. Res. Q.(1970-1977) **23**(3), 289–303 (1972)
4. Syntetos, A.A., Boylan, J.E.: On the bias of intermittent demand estimates. Int. J. Prod. Econ. **71**(1-3), 457–466 (2001)
5. Syntetos, A.A., Boylan, J.E.: The accuracy of intermittent demand estimates. Int. J. Forecast. **21**(2), 303–314 (2005)
6. Shale, E.A., Boylan, J.E., Johnston, F.R.: Forecasting for intermittent demand: the estimation of an unbiased average. J. Oper. Res. Soc. **57**(5), 588–592 (2006)
7. Teunter, R.H., Duncan, L.: Forecasting intermittent demand: a comparative study. J. Oper. Res. Soc. **60**(3), 321–329 (2009)
8. Steutel, F.W., van Harn, K.: Discrete analogues of self-decomposability and stability. Ann. Probab. **7**(5), 893–899 (1979)
9. McKenzie, E.: Some simple models for discrete variate time series. Water Resour. Bull. **21**, 645–650 (1985)
10. Al-Osh, M., Alzaid, A.: First-order integer-valued autoregressive (INAR(1)) processes. J. Time Ser. Anal. **8**, 261–275 (1987)
11. McKenzie, E.: Some ARMA models for dependent sequences of Poisson counts. Adv. Appl. Probab. **20**(4), 822–835 (1988)
12. Freeland, R.K., McCabe, B.P.M.: Forecasting discrete valued low count time series. Int. J. Forecast. **20**(3), 427–434 (2004)
13. Jazi, M.A., Jones, G., Lai, C.D.: First-order integer valued AR processes with zero inflated poisson innovations. J. Time Ser. Anal. **33**, 954–963 (2012)
14. Piancastelli, L.S.C., Barreto-Souza, W.: Inferential aspects of the zero-inflated Poisson INAR(1) process. Appl. Math. Model. **74**, 457–468 (2019)
15. Syntetos, A.A., Babai, M.Z., Dallery, Y., Teunter, R.: Periodic control of intermittent demand items: theory and empirical analysis. J. Oper. Res. Soc. (2008). https://doi.org/10.1057/palgrave.jors.2602593



Life Satisfaction and Institutional and Interpersonal Trust: A Comparative Analysis of Italian and Foreign Citizens

Martina Bodo[1(✉)], Angela Paparusso[1], and Elena Ambrosetti[2]

[1] Istituto di Ricerche Sulla Popolazione e le Politiche Sociali (IRPPS-CNR), Via Palestro, 32, 00185 Rome, Italy
martina.bodo@irpps.cnr.it

[2] Dipartimento MEMOTEF, Sapienza Università Di Roma, Via del Castro Laurenziano, 9, 00161 Rome, Italy

Abstract. Institutional trust and interpersonal trust are key ingredients of growth, individual and societal well-being, and governance. The aim of this paper is to study the subjective well-being (SWB) of the resident population in Italy, by analyzing the association between self-reported life satisfaction and institutional and interpersonal trust, controlling for sociodemographic, human capital and other measures of SWB, whilst distinguishing between Italian and foreign citizens. We used data from the cross-sectional survey on "Aspects of Daily Life", collected by the Italian National Institute of Statistics (ISTAT) from 2015–2021. The multivariate analysis is based on ordinary least squares regression (OLS) with wave fixed effects to account for changes in the distribution between different years. Results suggest that higher interpersonal trust is associated with higher levels of life satisfaction. Moreover, the association of institutional trust with life satisfaction is positive, albeit comparatively small in magnitude. Both Italian and foreign citizens show comparable outcomes.

Keywords: SWB · Institutional trust · Interpersonal trust · Citizenship

1 Introduction

To the best of our knowledge, studies focusing on immigrants' subjective well-being (SWB) are still very few in Italy. The limited number of studies on this topic [1–8] is mainly due to the scarcity of suitable and up-to-date survey data. In Italy, a comparison between immigrants and natives is complicated by the lack of a survey covering the topic of SWB for both populations. However, starting from 2015 the Multipurpose survey on households "Aspects of Daily Life" (Aspetti della Vita Quotidiana, AVQ), allows to distinguish between Italian and foreign citizens. We believe that information on the

This work was supported by the PRIN project 'The subjective well-being of immigrants and natives (WELL-MIG)' [grant number 2022CHTHEF], financed by the Italian Ministry for Education, Universities, and Research and the European Union – Nex Generation EU.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 208–214, 2025.
https://doi.org/10.1007/978-3-031-64431-3_35

respondents' foreign citizenship can be used as a proxy of their migration background. Examining disparities in SWB between Italian and foreign citizens is crucial for understanding if their well-being levels are similar and which factors might either foster or hinder parity.

Institutional trust and interpersonal trust are key ingredients of growth, individual and societal well-being, and governance [9]. Interpersonal trust encompasses the ability to trust others – including both strangers and known individuals –; whilst institutional trust refers to governments and public institutions fulfilling their role in a satisfactory and efficient manner according to the citizens' perceptions [10]. Both forms of trust might lead to feelings of certainty and security, and thus have a positive impact on life satisfaction. When trust is lacking it can lead to social unrest and economic instability, factors that are negatively associated with well-being.

Although previous studies have discovered a strong and positive correlation between SWB and institutional and interpersonal trust across European countries [11], such topic has been overlooked in Italy where, to the best of our knowledge, no studies have been conducted so far. Therefore, using the AVQ survey (2015–2021), the aim of this paper is to analyze the association between self-reported life satisfaction – a cognitive measure of subjective well-being – and variables of institutional and interpersonal trust, controlling for sociodemographic, human capital and other measures of subjective well-being, whilst distinguishing between Italian and foreign citizens.

In the next section, data, measures, and methods are presented, whilst the last section contains a discussion on the obtained results.

2 Data, Measures, and Methods

2.1 Data

We used data from the "Aspects of Daily Life" (AVQ), carried out annually by ISTAT since 1993 and providing valuable insight into the habits and the challenges faced by the resident population in their everyday life. For the purpose of our analysis, we pooled the data of the last seven waves covering 2015–2021. We restricted our sample to people aged 14 and over, which results in an initial sample of 278,729 individuals. The final sub-sample consists of 264,772 individuals (having excluded 13,957 individuals with missing information on one or more variables used in the statistical analyses).

2.2 Measures

Self-reported life satisfaction is our dependent variable, measured with a 11-point Likert scale from 0 (not satisfied at all) to 10 (very satisfied). The question was *"All things considered, how satisfied are you with your life as a whole nowadays?"*.

Institutional and interpersonal trust were measured with the following two questions from the survey: *"Using a score from 0 (not satisfied at all) to 10 (very satisfied), indicate how much you trust the following institutions: Italian Parliament, Regional government, political parties, judicial system, and law enforcement"* and *"Do you think you can trust most people or do you have to be very careful in dealing with people?"*.

We controlled for additional independent variables that have been demonstrated to be explanatory determinants of life satisfaction. Socio-demographic variables: (1) citizenship, (2) gender, (3) age in classes, (4) marital status. Human capital variables: (5) educational attainment, (6) occupational condition. Subjective well-being variables: (7) self-rated health, (8) environmental satisfaction, and (9) hope about the future.

2.3 Methods

As a first step, we carried out descriptive analyses. This is followed by multivariate analyses, based on linear regression models (OLS) with wave fixed effects to account for changes in the distribution between different years. We ran 3 separate models. The full model (1) includes all the individuals present in the sample. We then ran two separate models for (2) Italian and (3) foreign citizens to have a better understanding of differences in life satisfaction and associated factors between the two groups.

3 Results

3.1 Descriptive Results

Our respondents show a quite high self-reported life satisfaction (on average, equal to 7.047) with minor differences between Italian (7.05) and foreign (6.97) citizens during the period 2015–2021. Over the years, the performance disparity between Italian and foreign citizens has narrowed from 0.26 in 2015 to 0.04 in 2021.

Table 1 reports the descriptive statistics for all independent variables. The share of foreign citizens is stable across waves and is around 5%.

Table 1. Descriptive statistics of the independent variables. N = 264,772. Italy, 2015–2021. Source AVQ-ISTAT

Independent variables	% or mean	Independent variables	% or mean
Trust		Legally married	46.8
Institutional trust		Divorce/Widowed	18.1
Italian Parliament	4.0	**Human capital**	
Regional government	4.0	*Educational attainment*	
Political parties	2.8	Primary education or lower	17.3
Judicial system	4.5	Lower secondary education	30.2
Law enforcement	6.6	Upper secondary education	37.4
Interpersonal trust		Tertiary education	15.0
No	77.1	*Occupational condition*	
Yes	22.9	Employed	42.3

(*continued*)

Table 1. (*continued*)

Independent variables	% or mean	Independent variables	% or mean
Socio-demographic		Unemployed	10.3
Citizenship		Housewife	14.3
Italian citizens	95.2	Student	8.2
Foreign citizens	4.8	Other	25.0
Gender		**Subjective well-being**	
Male	47.8	*Self-reported health*	2.3
Female	52.2	*Environmental satisfaction*	2.2
Age		*Hope about the future*	
14–34	21.9	Will improve	27.5
35–54	33.4	Will stay the same	32.2
55–64	16.5	Will worsen	14.6
65+	28.1	I do not know	25.7
Marital status			
Single	31.3		

3.2 Multivariate Results

Table 2 displays the coefficients of the OLS regressions. Wave dummy variables were included in the regression but have not been reported. Our full regression model explains 23% of the total variance in self-reported life satisfaction.

Table 2. OLS regression models predicting self-reported life satisfaction of Italian and foreign citizens. Italy 2015–2021.

Independent variables	Model 1 N = 264,772	Model 2 N = 252,055	Model 3 N = 12,717
Trust			
Institutional trust			
Italian Parliament	0.01***	0.02***	–0.01
Regional government	0.04***	0.04***	0.05***
Political parties	–0.01***	–0.01***	0.01*
Judicial system	–0.01***	–0.01***	0.00
Law enforcement	0.11***	0.11***	0.09***
Interpersonal trust			

(*continued*)

Table 2. (*continued*)

Independent variables	Model 1 N = 264,772	Model 2 N = 252,055	Model 3 N = 12,717
No	–	–	–
Yes	0.16***	0.16***	0.13***
Socio-demographic			
Citizenship			
Italian citizens (reference)	–		
Foreign citizens	–0.21***		
Gender			
Female (reference)	–	–	–
Male	0.00	0.01	–0.07*
Age			
14–34 (reference)	–	–	–
35–54	–0.05***	−0.06***	−0.01
55–64	0.02	0.00	0.11*
65 +	0.24***	0.23***	0.32***
Marital status			
Single	–0.02*	−0.02*	0.07
Legally married	0.36***	0.36***	0.33***
Divorce/Widowed (reference)	–	–	–
Human capital			
Educational attainment			
Primary education or lower (reference)	–	–	–
Lower secondary education	0.10***	0.10***	0.09*
Upper secondary education	0.11***	0.10***	0.21***
Tertiary education	0.15***	0.14***	0.26***
Occupational condition			
Employed	0.08***	0.09***	0.11
Unemployed	−0.53***	−0.54***	−0.41***
Housewife	−0.06***	−0.06***	0.05
Student	0.07***	0.07***	0.16*
Other (reference)	–	–	–
Subjective well-being			

(*continued*)

Table 2. (*continued*)

Independent variables	Model 1 N = 264,772	Model 2 N = 252,055	Model 3 N = 12,717
Self-reported health	−0.45***	−0.46***	−0.37***
Environmental satisfaction	−0.30***	−0.30***	−0.33***
Hope about the future			
Will improve	0.47***	0.45***	0.65***
Will stay the same	0.20***	0.20***	0.21***
Will worsen	−0.73***	−0.73***	−0.66***
I do not know (reference)	–	–	–
R^2	0.23	0.23	0.22

After controlling for the abovementioned socio-demographic, human capital, and SWB variables, the coefficients of our chosen indicators of institutional trust are statistically significant, albeit comparatively small in magnitude (apart from law enforcement). On the other hand, interpersonal trust appears positive and a significant predictor of self-reported life satisfaction, for both Italian (model 2) and foreign citizens (model 3) – although having a foreign citizenship, on average, lowers life satisfaction (model 1). Such results are in line with previous findings: the coefficient of interpersonal trust tends to be stronger than that of institutional trust [12]. Moreover, a poor self-reported health and a negative environmental satisfaction are negatively associated with life satisfaction, while the opposite is true for being hopeful about the future.

We are aware of the limits of the present study. First, cross-sectional data do not allow to investigate causal links between trust and life satisfaction, likewise it was not possible to exclude reverse causality between life satisfaction and trust indicators and other SWB measures. Second, we are unable to study the factors related to the migration and integration process since the AVQ survey does not provide any information on this matter. Third, we acknowledge the fact that the category "Italian citizens" may also include naturalized citizens. However, we argue that the requirements and the lengthy process needed to obtain the Italian citizenship are sufficient to place birthright and naturalized citizens on the same level. Lastly, we are aware that Italian and foreign citizens might use different values and personal criteria when judging their levels of trust as well as their satisfaction with life. Nevertheless, for as long as there are no suitable and up-to-date survey data, we believe this study represents an important step towards advancing research on the SWB of Italian and foreign citizens in Italy.

References

1. Ambrosetti, E., Paparusso, A.: What are the main factors associated with immigrants' subjective well-being in Italy? evidence from self-reported life satisfaction. Int. Migr. **59**(4), 221–237 (2021)

2. Cela, E., Barbiano di Belgiojoso, E.: Ageing in a foreign country: Determinants of self-rated health among older migrants in Italy. J. Ethnic Migr. Stud. **47**(15), 3677–3699 (2021)
3. Paparusso, A.: Immigrant Integration in Europe. Springer, A subjective well-being perspective (2021)
4. Barbiano di Belgiojoso, E., Cela, E., Rimoldi, S.M.L.: The effect of migration experiences on wellbeing among ageing migrants in Italy. Soc. Indic. Res. **161**, 553–579 (2022)
5. Carella, M., García-Pereiro, T., Pace, R.: Subjective well-being, transnational families and social integration of married immigrants in Italy. Soc. Indic. Res. **161**(2–3), 785–816 (2022)
6. Tosi, F., Impicciatore, R.: Transnational parenthood and migrant subjective well-being in Italy. Int. Migr. Rev **56**(4), 1141–1166 (2022)
7. Ambrosetti, E., Bettin, G., Cela, E., Paparusso, A.: Subjective well-being and school outcomes among children of immigrants and natives in Italy. Popul. Space Place **29**(4), 1–15 (2023)
8. Campolo, M.G., Di Pino Incognito, A.: How perceived well-being determinants differ for immigrants and natives in Italy. Qual. Quant. 1–24 (2023)
9. OECD: OECD Guidelines on Measuring Trust, OECD Publishing, Paris (2017)
10. Sønderskov, K.M., Dinesen, P.T.: Trusting the state, trusting each other? the effect of institutional trust on social trust. Polit. Behav. **38**, 179–202 (2016)
11. Hudson, J.: Institutional trust and subjective well-being across the EU. Kyklos **59**(1), 43–62 (2006)
12. Jovanović, V.: Trust and subjective well-being: the case of Serbia. Pers. Individ. Differ. Individ. Differ. **98**, 284–288 (2016)

Population Change Across Italian Municipalities: Urbanisation and the Role of Connectivity

Danilo Bolano[1(✉)], Daniela Ghio[2], Alessandro Di Nallo[3], and Francesco Billari[1]

[1] Bocconi University, Milan, Italy
danilo.bolano@unibocconi.it
[2] Univerisità di Catania, Catania, Italy
[3] Max Planck Institute for Demographic Research (DE), Rostock, Germany

Abstract. This study investigates the Italian demographic landscape by analyzing urbanization, migration dynamics, and the impact of connectivity. Leveraging detailed Italian municipal-level data on population structure and migration patterns, and considering Italy's unique north-south demographic disparities, the research reveals distinct patterns in migration movements, especially emphasizing the growing appeal of non-urban settings. This study explores how improved connectivity bridges the urban-rural gap considering internet speeds as a key factor influencing migration dynamics. Our study contributes nuanced perspectives to migration studies and informs policy discussions, advocating for inclusive strategies that leverage technological advancements to foster balanced urban development and mitigate geographical disparities. It offers a robust foundation for understanding the intricate dynamics of migration, urbanization, and digital connectivity in modern societies.

Keywords: demography · population change · digital demography

1 Background

Populations experiencing advanced phases of demographic transition are also more likely to achieve high levels of urbanization [1,2]. A large variability has been detected across countries and over time [3]. Among the most disputed arguments is the role played by internal mobility and international migration as demographic components of population changes in rural areas and cities. The aim of the paper is to measure the contribution of international migration and internal mobility at the local level to empirically establish the role of these demographic components in shaping the population distribution across local administrative units (municipalities). The results from this analysis are expected to improve our understanding of the spatial configuration of urban and rural settlements, shedding light on the extent to which urbanization dynamics interplay

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 215–220, 2025.
https://doi.org/10.1007/978-3-031-64431-3_36

with population changes. In this vein, we also account for the disparity in connectivity as the contextual factor that, after the COVID-19 pandemic, has been recognized as a potential determinant of residential mobility [4].

2 The Italian Landscape

Italian society is characterised by strong family ties and a relatively weak social welfare system, which appear to favor immobility and discourage the Italian young generations from leaving parental home. Nevertheless, net migration of the 20 - 24 age group shows a negative balance in the Italian southern municipalities and a positive one in northern municipalities [5], but with a wide distribution across territories. In addition to an urban-rural divide, the demographic landscape in Italy has historically been marked by significant geographical imbalances. Despite the low fertility rate and a negative natural balance, most northern regions have exhibited population growth in the last two decades, driven by a positive migratory balance, while southern regions have faced a negative growth trend [6]. In this context, we aim to document the urban transition and population change across Italian macro-areas (corresponding to the European Union NUTS1 classification), moreover we will focus on both internal mobility and international migration [7,8].

3 Data and Methods

We use official figures provided by the ISTAT, which include age-specific population stocks and demographic components: births, deaths, migration, and mobility flows (both internal and international movements). These data are available at the municipal level and cover the period from 2011 to 2020. Data allow us to further distinguish between internal movements (i.e., movements within Italy, to/from another Italian municipality) and external movements (i.e., migrations to/from abroad). The Degree of Urbanisation classification established by the Eurostat has been employed to categorize municipalities into three classes: cities (densely populated areas where at least 50% of the population lives in urban centers, which we will refer to as urban areas), rural areas (thinly populated areas where more than 50% of the population lives in rural grid cells) or towns and suburbs (for simplicity we will refer to these as intermediate areas). Connectivity is assessed using internet download speeds sourced from Ookla's geolocated data, which reflects the actual Internet speeds experienced by users in each Italian municipality. We categorize municipalities based on their fixed internet network performance into three groups: i) low-speed internet (i.e., no broadband or download speeds up to 6 Mbps); ii) medium-speed broadband service (up to 24 Mbps); and high-speed internet (25 Mbps and over).

3.1 Measure of Population Change

To measure the population change and its components, we rely on a measure of population turnover discussed by [9]. For each municipality (k) in a specific year

(t), Population Turnover Rate (PTR) is defined as the algebraic sum of: crude annual birth rate (b), death rate (d), (internal) within-country in-mobility rate (i) (i.e., number of individuals who moved in the municipality k from another Italian municipality divided by total population), across border im-migration (external im-migration j) (i.e., considering those arrived in municipality k from abroad), (internal) within-country out-mobility (o) (i.e., those who moved out of municipality k to move to another Italian municipality), and across border out-migration (external out-migration m) (i.e., those who moved out of municipality k to go abroad).

While the numerical value of PTR cannot be interpreted per se, it must be seen in relative terms and it can be seen as the speed of population change [9]. Municipalities with a higher PTR have seen changes in their demographic landscape faster than municipalities with lower levels of PTR. Based on PTR, it is possible to calculate the so-called Migration Share of Turnover (MST):

$$MST_{k,t} = \frac{(i_{k,t} + j_{k,t}) + (o_{k,t} + m_{k,t})}{PTR_{k,t}}$$

The MST is bounded between 0 to 100 and represents the proportion of population turnover attributable to in-flow and out-flow population movements. In municipalities with an MST above 50, population movements have a greater impact than birth and death in explaining demographic change. We can further distinguish between MST internal where we consider only internal movements (i and o) and MST external where we focus solely on movements to and from abroad (j and k).

The analysis comprises two main steps. In the first phase, we describe the levels of PTR and MST municipalities over the last decade, by the degree of urbanization. Subsequently, we will examine the extent to which the PTR/MST of a municipality is associated with its level of connectivity and assess the heterogeneity of this relationship based on the degree of urbanisation. We will employ a local multiscale process (GWR geographically weighted regression) at the municipality level. Unlike conventional regression models, GWR allows us to explore potential spatial nonstationarity in relationships. We will investigate the relationship between connectivity and population change while considering the complex spatial in Italy, in terms of the north-south and urban-rural dynamics.

4 Preliminary Results

Urban areas show significantly lower levels of Population Turnover (see Fig. 1) concerning rural or intermediate municipalities with a PTR. In all three typologies, population movements account for more than 70% of the total population change with urban municipalities once again showing relatively lower levels. However, when distinguishing between internal and external population movements, we can clearly see two different trends. Urban cities are characterized by migratory movements to and from abroad (bottom left panel), while instead rural and intermediate areas see the predominant role of internal movements.

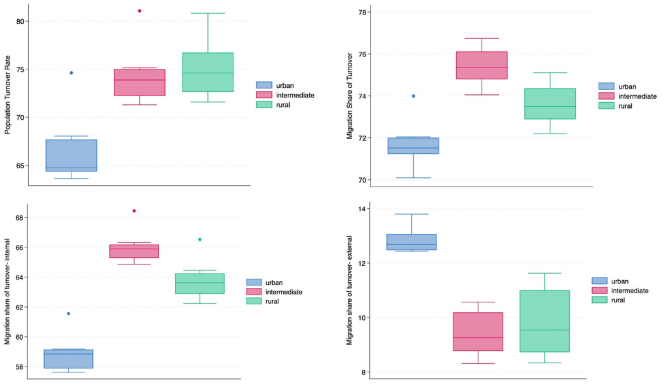


Fig. 1. Population Turnover and Migration Share by degree of urbanisation and macro-area (NUTS-1). Authors' elaboration based on ISTAT data. By external migration, we consider both in-migration from abroad and out-migration to abroad. By internal mobility, we consider both in-mobility from another Italian municipality and out-mobility towards another Italian municipality. Authors' elaboration based on ISTAT data.

The north-south differences in demographic patterns remain evident, even when accounting for the level of urbanization (Fig. 2). Particularly in rural and intermediate settings, the population change in the southern and island regions is significantly lower than that observed in the rest of the country. The north-south divide becomes even more pronounced when we focus on the role of external migration in urban municipalities. While moving to and from abroad accounts for less than 10% of the total population turnover in the south and islands, this value doubles in the northern regions. When examining internal movements, in the south we still observe relatively high levels of mobility from urban cities, likely toward other urban cities in the norther part of the country.

4.1 Internet Speed in Italy

Connectivity has seen a sharp improvement in Italy over the last few years (Fig. 3). In 2016 in more than 35% of the municipalities, residents had access to "slow" internet – defined as an average download speed below 6 Mbps, and in 70 not even a broadband (below 1 Mbps) – on the other hand, only 68 municipalities had a "fast" broadband. In 2019 however, only 21 municipalities remained with a "slow" level of connectivity with more than 30% having access to a fast, or very fast (above 100 Mbps) internet. Despite such relevant changes in the level of connectedness, a strong urban-rural divide remains. Nowadays (see Fig. 3 right panel), municipalities with no broadband or very low internet speed are only in the rural part of the country. Using spatial autoregressive models, we will leverage such temporal and geographical heterogeneity to estimate the contribution of the level of connectivity to (the speed of) demographic change (PTR levels) and migration patterns (MST).

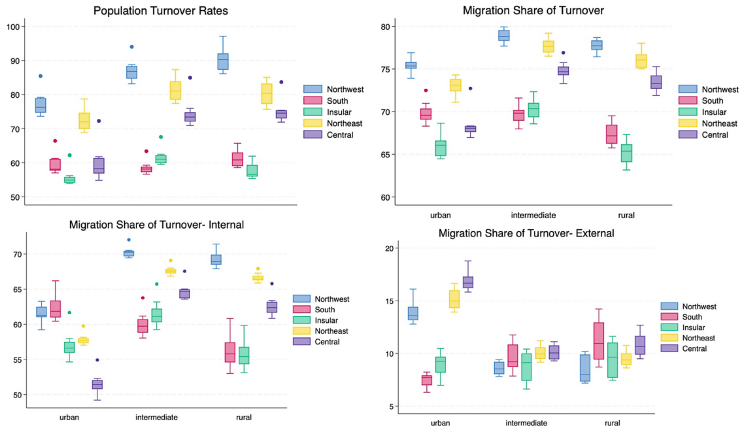


Fig. 2. Levels of Population Turnover and Migration Share by degree of urbanisation and macro-area (NUTS-1). By external migration, we consider both in-migration from abroad and out-migration to abroad. By internal mobility, we consider both in-mobility from another Italian municipality and out-mobility towards another Italian municipality. Authors' elaboration based on ISTAT data.

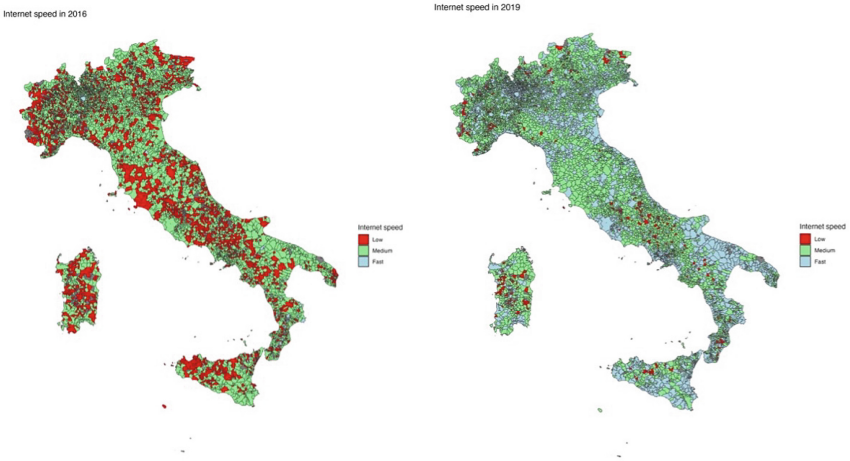


Fig. 3. Average Internet speed in Italian municipalities in 2016 (left panel) and 2019 (right panel). The color scheme represents different internet speeds with red indicating low speed, green indicating medium speed, and blue indicating fast speed.

5 Abridged Conclusions

In countries experiencing advanced demographic transition, mobility and migration can be seen as crucial factors in mitigating the effects of aging. Yet, the lack of data comparability and the absence of robust comparative metrics have posed significant challenges in understanding the spatial impact of migration

dynamics on population changes. We sought to address these gaps by utilizing official statistics on populations and migration flows at the municipality level in Italy. We will explore the role of connectivity, specifically internet speed availability, as a potential mechanism for reducing the gap between urban and rural areas, particularly in developed countries. Italy serves as a significant case study with several rural areas that still face challenges due to their remoteness from facilities and networks. Nevertheless, the well-established migration and mobility patterns of gains in the north and losses in the south of Italy may have been influenced by connectivity. Expanding connectivity to achieve an inclusive society has been at the heart of the policy agenda in all countries. Our study aims to contribute empirical evidence to the policy debate, shedding light on how connectivity can reshape spatial concentration and urban development.

Acknowledgements. The authors gratefully acknowledge the Italian National Institute of Statistics for the data provision. D. Bolano and A. Di Nallo acknowledge the financial support of FONDAZIONE CARIPLO Project n. 2021-1185 - FAST Population Change: Flows, AttractivenesS and Territories' resilience. (Pi. Bolano).

References

1. de Vries, J.: European Urbanization 1500-1800, Routledge, pp. 424 (2013)
2. Dyson, T.: The role of the demographic transition in the process of urbanization. Popul. Dev. Rev. **37**(Suppl 1), 34–54 (2011)
3. Jiang, L., O'Neill, B.C.: Determinants of urban growth during demographic and mobility transitions: evidence from India, Mexico, and the US. Popul. Dev. Rev. **2**, 363–389 (2018)
4. OECD: Bridging digital divides in G20 countries (2021)
5. Ghio, D., Bosco, C., Natale, F., Loeschmer, J., Goujon, A.: Age patterns of net migration and related urbanisation dynamics across European municipalities, Population Space and Place (2022)
6. Benassi, F., Busatta, A., Gallo, G., Stranges, M.: Le disuguaglianze tra territori. In ASIP - Rapporto sulla popolazione. L'Italia e le sfide della demografica (Billari FC e Tommasini C. Eds). Il Mulino, Bologna (2021)
7. Bonifazi, C. Mobile per forza. Spostamenti di popolazione nell'Italia della crisi. Il Mulino, 5, 798–805 (2013)
8. Impicciatore, R., Strozza, S.: Internal and international migration in Italy. An integrating approach based on administrative data. Polis **30**(2), 211–238 (2016)
9. Billari, F.C.: Demography: fast and slow. Popul. Dev. Rev. **48**(1), 9–30 (2022)



A Bayesian Two-Layer Latent Variable Model for Protein Inference

Jordy Bollon[1,2] and Simone Tiberi[3(✉)]

[1] Computational and Chemical Biology, Italian Institute of Technology, Aosta, Italy
[2] Astronomical Observatory of the Autonomous Region of the Aosta Valley, Nus, Italy
[3] Department of Statistical Sciences, University of Bologna, Bologna, Italy
Simone.Tiberi@unibo.it

Abstract. In this work, we study the problem of protein-isoform inference from mass spectrometry proteomics data. Inferring protein isoforms is complex because proteins are only indirectly measured via peptides. Here, we propose a Bayesian approach, based on a two-layer latent variable model to recover protein isoforms, starting from peptide-level data. Since isoform-level data is scarse, we further enhance information by embedding transcriptomics data (i.e., mRNA abundance), which are incorporated via informative priors. Our approach allows inferring the presence/absence and abundance of individual protein isoforms, and provides a measure of uncertainty of both estimates, via posterior probabilities and highest posterior density credible intervals. Notably, at present, existing proteomics tools do not allow inferring protein isoform abundances. Here, we present preliminary results, based on simulation studies, while more extensive benchmarks on real data are currently being performed. Our framework may be valuable for life scientists, and enable them to gain deeper insight into key biological mechanisms.

Keywords: Proteomics · Bayesian inference · latent variables · statistical software

1 Introduction

1.1 Background

Studying protein diversity and identifying physiologically relevant protein isoforms are crucial steps in biomedical research. Biological mechanisms, such as alternative splicing or alternative promoter usage, allow single genes to code for multiple isoforms [3]; for example, in humans, it was estimated that approximately 20,000 genes may give rise to over 300,000 isoforms [1,4]. At present, the main strategy to infer proteins is via "bottom-up" proteomics, where proteins are indirectly measured via peptides, which act as surrogate markers for their protein(s) of origin.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 221–225, 2025.
https://doi.org/10.1007/978-3-031-64431-3_37

However, inference on protein isoforms is affected by several technical limitations. In particular, most peptides, called shared peptides, map to multiple protein isoforms; furthermore, peptides may be erroneously detected, even when absent. As a consequence, most inferential approaches detect proteins at the gene level, and do not allow investigating proteome complexity at the isoform level.

1.2 Our Approach

Here, we describe a Bayesian approach, based on mass spectrometry proteomics data, to perform inference on individual protein isoforms. Our framework enhances the available data, by integrating transcript (i.e., mRNA) abundance, which is a prerequisite and correlate of protein abundance. In particular, transcriptomics data (e.g., RNA-sequencing) is employed to estimate the relative abundance (i.e., fraction) of transcript isoforms: these estimates are used to formulate an informative prior for the relative abundance of the respective protein isoforms. Note that, our method may also be used with proteomics data alone, but accuracy is expected to increase when transcriptomics data is also provided.

Since protein-level data is not observed directly, we employ a two-level latent variable model to recover it, starting from peptide-level data. In particular: first, we sample the presence/absence of each peptide based on its estimated probability of being mistakenly detected (provided directly by proteomics tools); second, only for shared peptides that were estimated as being present, we allocate their abundance across the protein isoforms they map to. This enables us to obtain the presence and abundance of each protein isoform. Additionally, we provide a measure of the uncertainty of our estimates, via the posterior probability that a protein isoform is present in the sample, and a 0.95 level highest posterior density credible interval of its abundance.

Importantly, to be best of our knowledge, at present, no tool infers protein abundance at the isoform level, and no method in proteomics has ever used a two-layer latent variable approach.

2 Results

2.1 Inference

Given P protein isoforms, we assume that the overall protein abundance of a sample, denoted by $n \in \mathbb{N}$, is distributed across the P isoforms according to a multinomial distribution:

$$X = (X_1, \ldots, X_P) \sim MN\left(\pi = (\pi_1, \ldots, \pi_P),\, n = \sum_{p=1}^{P} x_p\right), \tag{1}$$

where X_p represents the random variable indicating the overall abundance originating from the p-th protein (and x_p is its realization), and π_p is the probability that a unit of abundance comes from the p-th protein, with $\sum_{p=1}^{P} \pi_p = 1$. The probability that the p-th protein is present in the sample is estimated via $Pr\,(X_p > 0)$.

When mRNA data is available, we use an informative Dirichlet prior for π:

$$\pi \sim Dir\,(\delta = (\delta_1, \ldots, \delta_P))\,, \tag{2}$$

with δ proportional to the transcript isoforms abundances. If transcriptomics data is not available, we set $\delta_p = 1$ for $p = 1, \ldots, P$, which corresponds to a weakly informative prior. Since the Dirichlet is a conjugate prior for the multinomial distribution, this choice results in a convenient posterior distribution for π:

$$\pi | x, \delta \sim Dir((x_1 + \delta_1, \ldots, x_P + \delta_P)). \tag{3}$$

We further assume a weakly-informative discrete uniform prior for the overall abundance of proteins:

$$X_p \sim Unif(0, 1, \ldots, n), \text{ for } p = 1, \ldots, P. \tag{4}$$

2.2 Latent Variable Sampling

Although we aim to infer proteins, measurements refer to peptides; therefore, protein isoform abundances, in X, are unobserved and treated as latent variables, which are sampled via a data augmentation approach in two steps, as described below.

First, we sample if a peptide has been erroneously detected. Assume that N peptides are detected in total, and that PE_i is the estimated probability (taken as input) that the i-th peptide, albeit absent, is erroneously detected. We sample if a peptide has been wrongly detected following a Bernoulli distribution:

$$\epsilon_i | PE_i \sim Bern(PE_i), \text{ for } i = 1, \ldots, N, \tag{5}$$

where $\epsilon_i = 1$ if the $i-th$ peptide has been mistakenly detected, and 0 otherwise.

Second, only for peptides sampled as correctly detected, we spread their abundance to the protein(s) they are compatible with. We define Y_i as the abundance of the i-th peptide, and ψ_i as the list of proteins the i-th peptide maps to; we further denote by X_{pi} the (unknown) abundance of peptide i that comes from protein p. We assume that the abundance of the i-th peptide can be redistributed to the proteins in ψ_i according to a multinomial distribution:

$$(X_{1i}, \ldots, X_{Pi}) | \pi, \psi_i, Y_i, \epsilon_i \sim MN\left(\tilde{\pi}^{(i)}, Y_i(1 - \epsilon_i)\right), \text{ for } i = 1, \ldots, N, \tag{6}$$

where $\tilde{\pi}^{(i)} = \left(\tilde{\pi}_1^{(i)}, \ldots, \tilde{\pi}_P^{(i)}\right)$, with

$$\tilde{\pi}_p^{(i)} = \frac{\pi_p \mathbb{1}(p \in \psi_i)}{\sum_{p'=1}^{P} \pi_{p'} \mathbb{1}(p' \in \psi_i)}, \tag{7}$$

where $\mathbb{1}(A)$ is 1 is A is true, and 0 if A is false. In other words, $\tilde{\pi}_p^{(i)}$ is proportional to π_p if the i-th peptide maps to the p-th protein, and 0 otherwise. Note that, in (6), the peptide abundance, $Y_i(1-\epsilon_i)$, is 0 if the peptide has been sampled as mistakenly detected in (5) (i.e., when $\epsilon_i = 1$).

Finally, we recover the protein-isoform abundances by adding the abundances obtained from the N peptides allocations: $X_p = X_{p1} + \ldots X_{pN}$, for $p = 1, \ldots, P$.

Parameters and latent states are alternately sampled from their conditional distributions, following two Gibbs samplers, via a Markov chain Monte Carlo (MCMC) scheme. Convergence is assessed via Heidelberger and Welch's stationarity test.

2.3 Simulation Study Preliminary Results

We have designed a simulation study to assess the properties of our approach.

In order to simulate realistic patterns, we collected six real datasets (both RNA-seq and MS proteomics), and fit our method to infer the abundance of each protein isoform. Then, in every dataset, we used the real data peptide-to-protein connections, and allocated the estimated abundances of each protein isoform across all the peptides compatible with it; hence resulting in simulated peptide-level abundances. Finally, starting from each simulated dataset, we studied the ability of our method to perform inference on the original protein isoforms. To assess the robustness of our framework to the mRNA prior, we fit our method twice: with and without mRNA abundances. Performance was measured via three metrics: the area under the curve (AUC) for the protein isoform presence/absence (using the estimated probability of presence); the Pearson correlation between the real and the estimated protein isoform abundances (log-10 scale); the coverage of the 0.95 level highest posterior density credible interval for the protein isoform abundances.

Table 1. Results from the simulation study.

	Without mRNA			With mRNA		
Dataset	AUC	log10-corr	Coverage	AUC	log10-corr	Coverage
Data_1	0.91	0.85	0.98	0.97	0.95	0.99
Data_2	0.92	0.89	0.98	0.97	0.97	0.99
Data_3	0.87	0.83	0.98	0.96	0.95	0.98
Data_4	0.93	0.90	0.98	0.98	0.98	0.99
Data_5	0.95	0.92	0.98	0.98	0.98	0.98
Data_6	0.95	0.92	0.98	0.98	0.98	0.99

In all six simulations, we obtained satisfying AUCs and log10 correlations, and the coverage of the credible intervals is above 0.95; furthermore, results significantly improve when employing mRNA abundances as informative priors (Table 1).

3 Discussion

In this work, we have proposed a novel Bayesian method to perform inference on protein isoforms from mass spectrometry data; this task is challenging given that proteins are only indirectly measured via peptides. Here, we present a two-layer latent variable approach to recover protein isoform abundances. We further enhance the available information, by integrating proteomics and transcriptomics data, which is incorporated by means of an informative prior. We have validated our method in a simulation study, while more extensive benchmarks on real data are ongoing.

Our tool could be of utility to computational biologists, by unlocking an unexploited potential for biological discoveries. For instance, it was shown that proteins of the transcription factor MITF display changes in abundance (at the gene level) between melanoma subtypes [2]; our approach could allow estimating the presence and abundance of the respective individual protein isoforms, and investigating how they vary across cancer subtypes, hence enabling a deeper understanding of cancer driving mechanisms.

At present, our method infers parameters from single samples. In the near future, we will extend our model to jointly fit multiple samples (i.e., biological replicates) in a Bayesian hierarchical framework, hence allowing for sample-specific parameters, while sharing information across samples. This extended model will also enable differential testing between conditions (e.g., treated *vs.* untreated), hence identifying the individual protein isoforms that vary across conditions.

References

1. Deveson, W., et al.: Universal alternative splicing of noncoding exons. Cell Syst. **6**(2), 245–255 (2018)
2. Rabbie, R., Ferguson, P., Molina-Aguilar, C., Adams, D.J., Robles-Espinoza, C.D.: Melanoma subtypes: genomic profiles, prognostic molecular markers and therapeutic possibilities. J. Pathol. **247**(5), 539–551 (2019)
3. Tapial, J., et al.: An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. **27**(10), 1759–1768 (2017)
4. Willyard, C.: Expanded human gene tally reignites debate. Nature **558**(7710), 354–355 (2018)



Assessing Variables Importance When Clustering Enterprises

Ilaria Bombelli(✉), Alessio Guandalini, and Giorgia Sacco

Italian National Institute of Statistics, Rome, Italy
{ilaria.bombelli,alessio.guandalini,giorgia.sacco}@istat.it

Abstract. The motivation of the study lies in the fact that many Structural Business Statistics (SBSs) surveys must move from considering the Legal Unit (LU) as unit of interest towards considering the Enterprise (ENT) as such. Therefore, it may be necessary to study a different stratification of the sample for adapting and improving the usual sample design based on LU for facing this change. By applying K-prototype clustering algorithm, we were able to identify for several input dataset which variables influence the most the clustering partition. By considering the most-influential variables in the new definition of the strata, a different stratification can be implemented with the aim of reducing the dimensional complexity and preserving the efficiency of the estimates.

Keywords: Clustering · Enterprise · K-prototype · Stratification sampling design

1 Motivation of the Study

The EUROSTAT regulations, definitions and harmonization policies requires that the statistical units in the surveys related to Structural Business Statistics (SBS) must be defined in the same way in all the Member States to make comparable estimates.

Therefore, an interesting and challenging task which all European National Statistical Institutes (NSI) are facing is the transition from considering the legal unit (LU) as the statistical unit of analysis towards considering the Enterprise (ENT) as such. The LU is defined as *an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place,*[1] while the ENT is defined as *the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit.*[2]

[1] EUROSTAT glossary accessible at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Local_unit.

[2] Regulation (CEE) n. 696/93 accessible at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31993R0696.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 226–232, 2025.
https://doi.org/10.1007/978-3-031-64431-3_38

It is important to mention that each LU can belong to more than one ENT simultaneously according to a given *share*.

Clearly, the aforementioned transition is not trivial, as many NSI still need to provide estimates at a LU-level, for comparability through the time. For this reason, the intermediate step will be to consider the ENT as a statistical unit during the sampling process, but then provide the questionnaire to all the LUs belonging to the sampled ENT (according to the cluster sampling strategy). In this way, data will be collected at a LU-level, but the estimates are possible at both levels, by properly adjusting the sampling weights in the calibration step. This paper aims to provide some suggestions for leading this transition process. In particular, we aim to bring out some hints for adapting the sampling design to this change and for stratifying the ENT in an efficient way.

The data and the methodological tools descriptions are provided in Sect. 2, 3, respectively. Results are presented and discussed in Sect. 4. Finally, Sect. 5 sums up the main outcome of the research.

2 Data

Data collecting information about the ENT are all reported into the Business Register ASIA, that includes all enterprises carrying on economic activities contributing to gross domestic product at market prices, in the fields of industry, commerce and services.[3]

Our analysis is focused on considering the most updated version of the register available, i.e. ASIA ENT 2021, the register including the ENT active in 2021. The dimension of the whole dataset is 4462146×17.

We decided to focus our study on different subsets of ASIA ENT 2021, obtained by filtering on the variable related to the *statistical classification of economic activities in the European Community* (NACE): a) ENT working in the C section of the NACE classification ('manufacturing activities'); b) ENT working in the M section ('Professional, Scientific and Technical Activities'); c) ENT working in the R section ('arts, sports, entertainment and recreation activities') ; d) ENT working in the division codes (NACE 2-digits) usually survey in Community Innovation Survey (CIS).

In addition, we consider only a subset of the available variables, which we noticed to be of interest in the stratification step. In particular we considered the variables *number_ug_ent*, *employees_ent*, *turnover*[4], *ls_ent*, *region_ent*, *ateco5_ent*. In addition, we further manipulated some of them to obtain additional variables such as *s_shares*, *s_shares_diff*, *ateco4_ent*, *ateco3_ent*, *ateco2_ent*, *section_ent* and *cladd3_ent*. The *ateco* variables are related to the ENT's economic activity and

[3] For a complete description of the register the reader may refer to Siqual website accessible at: https://siqual.istat.it/SIQual/visualizza.do?id=7777976.

[4] In order handle missing data in this variable, we used the Multivariate Imputation by Chained Equation (mice `R` package, van Buuren and Groothuis-Oudshoorn [2011]), even though we are aware that there exist many other imputation techniques which are used in Official Statistics.

follow the NACE classification. The variables' description and information are summarized and reported in Table 1.

Table 1. Variables' description

Variable	Description	Type
number_ug_ent	number of UG belonging to the ENT	numeric
employees_ent	number of employees working for the ENT	numeric
turnover	turnover	numeric
s_shares	sum of the shares of the LU belonging to the ENT and sharing with it the same NACE code	numeric
s_shares_diff	sum of the shares of the LU belonging to the ENT and with a NACE code different from the ENT's code	numeric
ls_ent	legal status	categorical
region_ent	region	categorical
ateco5_ent	NACE code (5 digits)	categorical
ateco4_ent	NACE code (4 digits)	categorical
ateco3_ent	NACE code (3 digits)	categorical
ateco2_ent	NACE code (2 digits)	categorical
section_ent	NACE code (1 digit)	categorical
cladd3_ent (cis)	class of employees: 10-49, 50-249, 250+	categorical

3 Methodological Tools

We applied cluster analysis in order to detect clusters of ENT and to assess which variables most influence the obtained cluster partitioning.

Since our dataset is a mixed data type, meaning that it includes both numerical and categorical variables, we applied the K-prototype clustering algorithm [Huang, 1998] which is an extension of K-Means [McQueen, 1967] and K-Modes [Huang, 1997] algorithm. It allows both numerical and categorical variables and it is implemented in the `clustMixType` `R` package [Szepannek, 2018]

For handling huge data dimension, instead of applying cluster analysis on the whole dataset, we used a bootstrap-like procedure: we selected 500 random samples of 1000 units without replacement from the entire dataset and on each of them we performed the cluster analysis. In particular, we ran K-prototype on each sample for different number of clusters varying in $[2, 20]$: the optimal number of clusters has been selected according to the Silhouette Method [Rousseeuw, 1987]. Once the number of clusters k^* has been chosen, we ran K-prototype on each sample to obtain k^* clusters. Then, we measured the feature importance of each variable by using the `R` function `FeatureImpCluster` [Pfaffel, 2021] and

we stored their ranking. The function `FeatureImpCluster` measures the importance of a feature (variable) by the misclassification rate relative to the baseline cluster assignment due to a random permutation of feature values. Therefore, the Misclassification Rate is considered as a measure of Feature Importance and the higher the Misclassification Rate, the greater the importance.

This study turns out to be very useful when dealing with high-dimensional datasets, as in our case. However, in order to validate the results, we also decided to implement K-prototype on the entire dataset of CIS survey to assess which variables were the most important and can be considered for stratification.

4 Results

Table 2 reports the variables which resulted to be the most influencing in the most of the iterations of the study separately for each dataset.

Table 2. Two most influencing variables in the simulation study. Datasets C, M, R refer to the subsets of ENT of ASIA2021 working in section C, M, R of NACE classification, respectively.

Dataset	Two most influencing variables	% of occurrences
C	ls_ent, s_shares	82.2, 6.0
M	region_ent, ateco5_ent	74.6, 21.8
R	ateco5_ent, region_ent	60.0, 39.4
univ CIS	section_ent, region_ent	92.0, 8.0

It highlights that in the dataset related to the CIS survey, the two most influencing variables are the section, which refers to the economic macro-activity, and the region.

For what concerns the two economic sector identified by section M and R, we noticed that the location information is very important: indeed, the location where the enterprises are placed may influence the economic activity of them. In addition, given that those two sectors are very heterogeneous, including a large number of quite different economic sub-activities, also the variable *ateco5_ent*, which refers to a very detailed economic activity description (being the most fine economic activity classification) turns out to be a very discriminant variable. Therefore, for these sectors, the location of the enterprises and the specific economic sector each enterprise works in, are the variables whose influences in discriminating the ENT are the highest.

Finally, for what concerns the ENT having economic activity represented by the section C the most influencing variable is the legal status.

In order to properly validate the results of the aforementioned study, we also applied the K-prototype to the entire dataset corresponding to the overall universe of CIS. The results show that the whole universe is split into 18 clusters

and that the most influencing variables are *region_ent* and *section_ent* which are consistent and give strength to the simulation's output. In order to further investigate the results, we decided to split the dataset into two subsets: the former considers only the ENT made up by one LU (*ENT-Mono*), the latter considers only the ENT made up by more than one LU (*ENT-Pluri).* Indeed, we believed that those two kind of ENT have very different structures and peculiarities. We applied K-prototype algorithm to both the subsets. The detailed variable's importance plots for the two subsets are shown in Fig. 1. In particular, for a given feature, the Misclassification Rate distribution is shown.



(a) ENT-Mono (b) ENT-Pluri

Fig. 1. Feature's importance when clustering ENT belonging to CIS universe

For *ENT-Mono*, the two most important variables are *section_ent* and *region_ent*, as it is highlighted also in the simulation study. Instead, for the *ENT-Pluri*, which is a very tiny subset of the whole dataset, the variables scoring high in terms of importance are *section_ent*, *ateco2_ent*, *cladd3_ent* and *region_ent*. Therefore, what is highlighted here is the possibility to obtain a new stratification which is more precise and detailed when the ENT is a ENT-Pluri, and more wide and comprehensive when the ENT is a ENT-mono.

Hence, to conclude the experiment, we decided to test the new proposed stratification where for each ENT i the stratum variable is defined as follows:

$$\text{stratum}_i = \begin{cases} \text{region}_i \times \text{section}_i & \text{if the ENT i is a ENT-mono} \\ \text{region}_i \times \text{ateco2}_i \times \text{cladd3}_i & \text{if the ENT i is a ENT-pluri} \end{cases}$$

Notice that for the ENT-pluri, we decided to exclude section variable, as we include ateco2 which is more accurate in describing the economic activity (being a subset of section).

The proposed stratification has been used in order to allocate a sample of the same dimension (39000 LU) of the one extracted with the current stratification. The number of new strata is reduced by around 800 with respect to the current stratification. Through a Monte Carlo simulation ($Nrep = 100$), we derived the estimates of *Employees* and *Turnover* variables and their related coefficient of variation (CV). The CV of an estimate is defined as the ratio of the standard deviation to the mean and is a well-known measure of the estimate's error widely used in many NSI. A common practice is to publish estimates with a CV lower

than 0.33 and consider fully reliable those with a CV lower than 0.15[5]. The median values of the CVs of the estimates resulting from the samples extracted according to the proposed allocation are slightly above the ones related to the current allocation, but turn out to be below the acceptance threshold. Details are provided in Table 3.

Table 3. Median values of estimates' CVs resulting from the samples selected either using the standard allocation or the new one.

DOM	Variable	CV_alloc_stand	CV_alloc_new
DOM1	Employees	0.017	0.064
DOM2	Employees	0.016	0.013
DOM3	Employees	0.015	0.009
DOM4	Employees	0.013	0.035
DOM1	Turnover	0.037	0.110
DOM2	Turnover	0.045	0.066
DOM3	Turnover	0.028	0.022
DOM4	Turnover	0.024	0.057

5 Final Consideration

In this work we decided to evaluate the effectiveness of a new stratification procedure in order to select a sample of ENT. The stratification variables have been identified thanks to K-prototype clustering method applied on mixed-data type such as the ASIA ENT register. The proposed stratification is used in order to allocate a sample of the same dimension of the one extracted with the current stratification. From the sample, we estimate some of the survey's target variables and their coefficient of variation (CV). The CVs are compared with the ones resulting from the current stratification. The comparison reveals that the efficiency of the estimates is preserved. In addition, the new stratification allows reducing the number of strata and therefore also the processing time is limited. However, further investigation are needed.

References

Huang, Z.: A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Dmkd 3.8, pp. 34–39 (1997)

Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. In: Data mining and knowledge discovery 2.3, pp. 283–304 (1998)

[5] See Statistics Canada. Sampling error, https://www150.statcan.gc.ca/n1/pub/62f0026m/2010003/section1-eng.htm.

McQueen, J.B.: Some methods of classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)

Pfaffel, O.: FeatureImpCluster: Feature Importance for Partitional Clustering. R package version 0.1.5. URL: https://CRAN.R-project.org/package=FeatureImpCluster (2021)

Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)

Szepannek, G.: ClustMixType: user-friendly clustering of mixed-type data in R. In: R J. 10.2, p. 200 (2018)

van Buuren, S., Groothuis-Oudshoorn, K.: mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45.3, pp. 1–67 (2011). https://doi.org/10.18637/jss.v045.i03



A Longitudinal Study of the Gender Gap in School Grades via Flexible Bayesian Beta Regression

Laura Bondi[1], Beatrice Franzolini[2], and Marco Palma[3](✉)

[1] Health Data Science Centre, Human Technopole, Milan, Italy
[2] Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy
[3] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
marco.palma@mrc-bsu.cam.ac.uk

Abstract. The gender disparity in school grades, known as the grade gender gap, has become a focal point for researchers, educators, and policymakers. Despite this heightened interest, there remains a scarcity of longitudinal studies addressing this issue. This study seeks to address this gap by examining the grade gender gap in schools using longitudinal data from Rimini, Italy. A Bayesian Beta regression with splines is devised to detect gender differences across different cohorts of students and along the educational journey from primary to upper-secondary school. Our findings reveal a significant gap in favour of female students, which widens during secondary school years but diminishes in younger cohorts. These pages contain preliminary results of the extended study in [2].

Keywords: Bayesian distributional regression · gender differences · panel data · regression splines

1 Gender Gap in School Grades

The *grade gender gap* is the difference in school performance between male and female students across various subjects and educational levels and is often used as a key metric by researchers, educators, and policymakers in recent years [6,8]. Quantification and investigation of this disparity may underscore potential imbalances within educational systems and their implications for individuals and society. While gender differences in school performance are a flourishing area of research, many studies are based on cross-sectional data and therefore do not capture changes over time. The aim of this study is to quantify and analyse the grade gender gap within primary and secondary schools over multiple years, analysing longitudinal data from the province of Rimini, Italy.

Recent studies conducted in Western nations indicate that, on average, girls tend to outperform boys academically, with narrower advantages in STEM subjects [see, for instance, 10,11]. However, gender disparities in favour of female students in STEM are not universally consistent. For example, findings from the 2022 OECD PISA report [9] reveal slight yet significant differences favouring boys in mathematics and science, despite girls exhibiting superior reading abilities. Notably, the literature highlights the

MRC Biostatistics Unit, University of Cambridge
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 233–238, 2025.
https://doi.org/10.1007/978-3-031-64431-3_39

fluctuations of the gender gap across historical epochs and geographical regions. Such variability highlights the influence of cultural, social, and economic factors on the magnitude and direction of this gap, thereby guiding interventions and policies aimed at its mitigation. These nuances display the complexity of the grade gender gap and stress the need for further research and monitoring of the phenomenon.

2 Data Description and RiminInRete Database

The Italian pre-university education system comprises three main cycles of studies: primary school (PR, five years), lower-secondary school (S1, three years), and upper-secondary school (S2, five years). Prior to primary school, children often attend three years of non-compulsory nursery school, and after upper-secondary school around 60% of students enrol in university [1]. The curriculum is consistent across all primary and lower-secondary schools, while in upper-secondary schools there is a broader array of tracks available. These are typically categorized into high schools (i.e., *licei*), which offer a general education primarily aimed at preparing students for subsequent university studies, and vocational schools (i.e., *istituti tecnici* and *istituti professionali*), which supplement standard topics with specialized training in technical and professional fields. In the Italian grading system, student performance is evaluated on a scale from 0 to 10, with 10 being the highest achievable grade. Grades below 6 are considered insufficient (fail), while those above 6 indicate varying levels of proficiency and success in each discipline. In primary schools, this grading system (with the regulatory framework of OM 172/2020) recently underwent a substantial change from numerical grades to a judgement-based system.

The data analyzed in this work are a stratified random sample from the RiminiInRete (RIR) database (https://www.rimininrete.net/). RIR is an initiative established in 2015 by a network of schools in the province of Rimini, Italy, to create a comprehensive database of periodic and final assessments across all schools in the province, to assess the efficacy of the evaluation system and to monitor the influence of assessment methodologies on students' academic advancement. The longitudinal nature of the RIR database allows for the analysis of the grade gender gap and its comparison among different student cohorts and stages of the educational journey.

We select $n = 1920$ students from the RIR database using a stratified sample by gender (as recorded by the school from the ID documents) and year of birth (from 1997 to 2012) with $n_s = 60$ students per stratum. A summary of the stratified sample is reported in Table 1. We consider their grade point averages (GPAs) over five consecutive academic years from a.y. 2014/2015 to a.y. 2018/2019 for a total number of observations equal to 6351. It is worth noting that the years considered in our analysis are not affected by the recent change in the grading system.

3 Flexible Bayesian Beta Regression Model

Beta regression [4,7] is a regression model tailored for response variables within the interval $(0,1)$. Unlike standard linear regression models (which work for unbounded continuous outcomes), Beta regression provides a coherent support for the response

Table 1. Average GPA, standard deviation, and effect size (Cohen's d) by gender, academic cycle, and academic year in the stratified sample of 1920 students.

	Number of observations		Mean		Stand. dev.		Cohen's d
	Female	Male	Female	Male	Female	Male	
Primary (PR)	1165	1180	8.7618	8.4822	0.8477	0.8828	0.3231
Lower-secondary (S1)	784	775	7.9112	7.3964	1.0087	0.9320	0.5301
Upper-secondary (S2)	1202	1209	7.3721	6.9761	0.9565	0.9731	0.4103
a.y. 2014/15	512	501	7.9798	7.6038	1.1549	1.0783	0.3364
a.y. 2015/16	655	667	8.0326	7.6301	1.1112	1.1888	0.3498
a.y. 2016/17	631	645	8.0717	7.6413	1.1353	1.1759	0.3722
a.y. 2017/18	669	664	7.9861	7.6102	1.1026	1.1712	0.3305
a.y. 2018/19	684	687	8.0236	7.7069	1.0620	1.0958	0.2935

variable and appropriately handles skewness in errors. It can be easily generalized to responses in a bounded interval (a,b) via data transformations [5].

Denoting with $y_{i,t}$ the GPA of student i for the academic year t rescaled by 10.05 to fit into the interval $(0,1)$, we use the following Bayesian model:

$$\begin{aligned} y_{i,t} \mid \mu_{i,t}, \phi_i &\sim \text{Beta}(\mu_{i,t}\phi_i, (1-\mu_{i,t})\phi_i) \\ \text{logit}(\mu_{i,t}) &= \alpha^{(\mu)} + \beta^{(\mu)} x_i + \gamma_{x_i} \times t + g_{x_i}(z_{i,t}) + b_i^{(\mu)} \\ \log(\phi_i) &= \alpha^{(\phi)} + \beta^{(\phi)} x_i + b_i^{(\phi)} \end{aligned} \tag{1}$$

with

$$\begin{bmatrix} b_i^{(\mu)} \\ b_i^{(\phi)} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^{(\mu)^2} & 0 \\ 0 & \tau^{(\phi)^2} \end{bmatrix} \right) \tag{2}$$

so that $\mathbb{E}[y_{i,t}] = \mu_{i,t} \in (0,1)$ and $\text{Var}[y_{i,t}] = \frac{\mu_{i,t}(1-\mu_{i,t})}{\phi_i+1}$. The regressors consider for the mean-submodel are the *gender* (x_i) of the student, the *academic year (t)* of the recorded GPA - whose regression coefficient (γ_{x_i}) is allowed to change based on the gender of the student - and a non linear effect g of the school level $z_{i,t}$ attended by the student at time t (from the first year of primary school to the fifth year of upper-secondary school) which is also allowed to change based on gender. The smooth function g is estimated via thin plate regression splines [12]. We assume the positive precision parameter ϕ_i to be time-independent and gender-dependent (through the regression coefficient $\beta^{(\phi)}$). Finally, we incorporate two student-specific random effects - one for the mean, $b_i^{(\mu)}$, and another for the precision, $b_i^{(\phi)}$ - to take into account unobserved individual-specific characteristics.

Priors for population-level linear effects are non-informative improper flat prior over the real numbers. Priors for the standard deviations of the random effects and splines are weakly-informative Half Student-t prior with 3 degrees of freedom. See [3] for more details on the prior choice.

The model was fitted using the R package `brms` [3] for Bayesian multilevel distributional regression. We run 2 chains of 3000 iterations each (of which 2000 warm-up).

4 Results

Table 2 and Fig. 1 summarise the results. Our analysis reveals several notable trends in the academic performance. Firstly, females demonstrate a consistent advantage, outperforming males on average throughout the entirety of their school careers. In particular, for the a.y. 2014/2015, the expected GPA for a female student is 0.4503 points higher than for her male peers. Nonetheless, the signs of the gender-dependent coefficients γ_F and γ_M associated with time suggest a gradual reduction on average in the gender gap in more recent years. This led to an average grade gender gap of 0.3288 points in the year 2018/2019. Refer to Table 3 and Fig. 1 for details on the temporal trend across the considered academic years.

Interestingly, we identify a significant positive coefficient for the scale parameter associated with male students, indicating that GPA tends to be less variable among male students compared to their female peers.

Figure 1, which includes both linear and non linear effects, showcases the estimated dynamic of the grade gender gap (represented in Fig. 1 by the distance between the two curves) for different school grade levels. While changes in the gender gap across different school grade levels are relatively limited for any fixed academic year, the gap appears to undergo a slight increase during lower and upper secondary schools. For both males and females, we also note that grades exhibit a relevant drop from the end of one school cycle to the start of the next, which is followed by a small rebound (possibly stronger for females on average) both at lower and upper secondary school.

Table 2. Posterior estimates for the coefficients in the Beta regression model in 1. l-95% CI and u-95% CI indicate the lower and upper bound of the 95% credible interval for each coefficient, respectively.

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
μ submodel (logit scale)					
Intercept ($\alpha^{(\mu)}$)	1.4789	0.0239	1.4320	1.5262	1.01
SexM ($\beta^{(\mu)}$)	-0.2729	0.0348	-0.3391	-0.2000	1.01
SexF:a.y (γ_F)	-0.0073	0.0059	-0.0191	0.0041	1.01
SexM:a.y (γ_M)	0.0109	0.0054	0.0007	0.0213	1.01
ϕ submodel (log scale)					
ϕ_Intercept ($\alpha^{(\phi)}$)	4.4809	0.0537	4.3746	4.5870	1.00
ϕ_SexM ($\beta^{(\phi)}$)	0.2409	0.0737	0.0993	0.3861	1.00
SD of random effects					
SD(Intercept) ($\tau^{(\mu)}$)	0.5484	0.0125	0.5242	0.5734	1.01
SD(ϕ_Intercept) ($\tau^{(\phi)}$)	1.0504	0.0388	0.9754	1.1276	1.00

Table 3. Grade gender gap estimates by academic year. Top row: posterior mean. Bottom row: 95% credible interval.

a.y. 2014/2015	a.y. 2015/2016	a.y. 2016/2017	a.y 2017/2018	a.y. 2018/2019
0.4503	0.4199	0.3894	0.3591	0.3288
[0.3299, 0.5576]	[0.3104, 0.5129]	[0.2868, 0.4755]	[0.2641, 0.4453]	[0.2297, 0.4199]

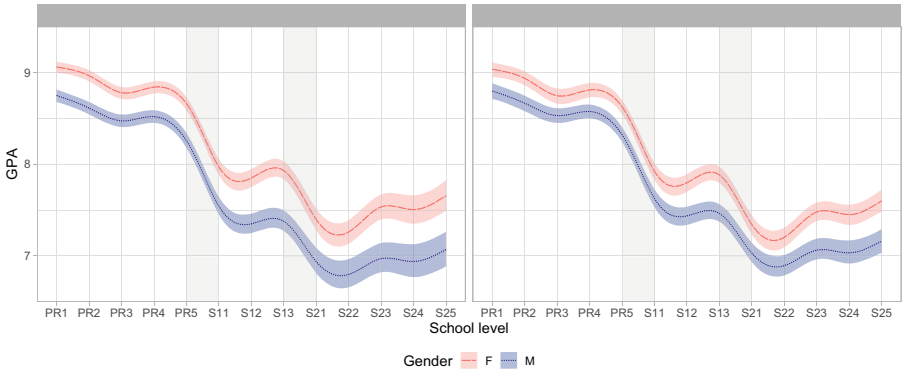


Fig. 1. Conditional effect of the school level by gender on GPA for two selected academic years. The areas in gray represent the change from one school cycle to the next.

5 Conclusions and Further Developments

This work highlights the existence of differences along the entire academic trajectories in a random sample of male and female students enrolled in schools of the province of Rimini between the academic years 2014/2015 and 2018/2019. Females outperform males on average across all the school levels, but the gender gap decreased on average in more recent years. To further investigate this gender gap, future work may focus on quantifying the discrepancy between performances within specific disciplines, e.g. gender gap in humanities or maths, with particular focus on gender-related changes from one school cycle to the next.

Another research direction is aimed at evaluating the impact on school grades of the COVID-19 pandemic and the subsequent disruption of in-person school attendance. To this aim, the plan is to extend the analysis to include data collected from academic years 2019/2020 to 2022/2023, which are now available from the RIR database.

Acknowledgements. We would like to acknowledge RiminInRete https://www.rimininrete.net/ for providing the data. Franzolini was supported by the National Recovery and Resilience Plan of Italy (PE1 FAIR - CUP B43C22000800006). Palma was supported by the UK Medical Research Council (MR/V020595/1).

References

1. AlmaLaurea: Rapporto 2020 sulla condizione occupazionale e formativa dei diplomati (2020)
2. Bondi, L., Franzolini, B., Palma, M.: A Bayesian longitudinal study of the grade gender gap. Working paper (2024)
3. Bürkner, P.C.: brms: an R package for Bayesian multilevel models using Stan. J. Stat. Softw. **80**, 1–28 (2017)
4. Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. J. Appl. Stat. **31**(7), 799–815 (2004)
5. Geissinger, E.A., Khoo, C.L., Richmond, I.C., Faulkner, S.J., Schneider, D.C.: A case for beta regression in the natural sciences. Ecosphere **13**(2), e3940 (2022)
6. Goodier, M., Aguilar García, C., Kirk, A., Scruton, P.: Gender gap shrinks and regional gap widens: 2023's key GCSE trends in England. The Guardian (2023)
7. Kieschnick, R., McCullough, B.D.: Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Stat. Model. **3**(3), 193–213 (2003)
8. Malespina, A., Singh, C.: Gender gaps in grades versus grade penalties: why grade anomalies may be more detrimental for women aspiring for careers in biological sciences. Int. J. STEM Educ. **10**(1), 13 (2023)
9. OECD: PISA 2022 results (Volume I): The state of learning and equity in education (2023)
10. O'Dea, R.E., Lagisz, M., Jennions, M.D., Nakagawa, S.: Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. Nat. Commun. **9**(1), 3777 (2018)
11. Voyer, D., Voyer, S.D.: Gender differences in scholastic achievement: a meta-analysis. Psychol. Bull. **140**(4), 1174 (2014)
12. Wood, S.N.: Thin plate regression splines. J. R. Stat. Soc. Ser. B Stat Methodol. **65**(1), 95–114 (2003)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Confidence Interval for the Complexity Index of Functional Data

Enea G. Bongiorno(✉), Lax Chan, and Aldo Goia

Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, Via Perrone 18, 28100 Novara, Italy
{enea.bongiorno,lax.chan,aldo.goia}@uniupo.it

Abstract. This paper presents the concept of complexity index for generic random processes. This is based on the notion of small-ball probability and the possibility that this probability can be decomposed as a factorization of two functions. One of these two factors carries information regarding the complexity of the process, which coincides with the number of random sources characterizing the process or, equivalently, its degrees of freedom. This factor has been studied in literature and this paper shows some statistical results on how it can be exploited to make inference on the degrees of freedom of the random process.

Keywords: Small-ball probability · 1–approximating functional · β–mixing · degrees of freedom

1 Introduction

Functional statistics deals with the study of processes whose observations are curves, surfaces or functions defined on a continuum (see [4–6,8]). In this area, a recently developed research topic study the complexity of a process by exploiting the notion of small-ball probability. In particular, given a random element X valued in a suitable metric space $\mathcal{F}$ and denoting by $B(\chi, h)$ the ball centered at $\chi \in \mathcal{F}$ with radius $h > 0$, the small ball probability of X is $\mathbb{P}(X \in B(\chi, h))$ when h tends to zero. Often it is convenient to assume that

$$\mathbb{P}(X \in B(\chi, h)) \sim \psi(\chi)\,\phi(h) \qquad \text{as } h \to 0, \tag{1}$$

with the identifiability condition $\mathbb{E}[\psi(X)] = 1$. Factorization (1) is not a forcing assumption; indeed it holds true under appropriate hypothesis (see, for instance, [1] and references therein). Notably, the function $\psi(\chi)$ can be interpreted as a surrogate density of the process X whereas the function $\phi(h)$ plays the role of the volumetric term and can be used to measure the complexity of the probability law of the process X (see [2]).To better understand this last idea, note that for some special family $\phi(h)$ exhibits a parametric form. For example, if the process is characterized by a finite number of independent random sources (as in the case of finite dimensional processes, see [3]) then $\phi_\theta(h) = c_\theta h^\theta$, for some constant term

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 239–242, 2025.
https://doi.org/10.1007/978-3-031-64431-3_40

c_θ and θ being the number of characterizing random sources or, equivalently, the degree of freedom of X. Another notable example comes from some infinite dimensional Gaussian processes (see [7]), for which $\phi_\theta(h) = C_1 h^\alpha \exp\left\{-C_2/h^\beta\right\}$ with $\theta = (\alpha, \beta) \in [0, \infty) \times (0, \infty)$ and positive constants C_1, C_2. In other words, the form of the volumetric term provides information related to the complexity of the probability law of the process and the parameter θ can be interpreted as a measure of the complexity.

This paper presents some recent efforts [2,3] in studying the complexity factor ϕ and complexity parameter θ. In particular, Sect. 2 presents a nonparametric estimator of ϕ that is then used to define the complexity index estimator based on the cosine dissimilarity measure. In Sect. 3, some asymptotics for the estimate of the complexity index are collected and in Sect. 4 an asymptotic confidence interval for that parameter is derived.

2 Estimation of the Complexity

The complexity factor $\phi(h)$ can be estimated by using its empirical version defined as follows. Given a sample of n curves $X_1, \ldots, X_n$ drawn from X, and $h > 0$ sufficiently close to zero, the empirical estimator of ϕ is

$$\widehat{\phi}_{emp}(h) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i \neq j} 1_{B(X_j,h)}(X_i), \tag{2}$$

proposed as instance in [3], where $1_A(x)$ is the characteristic function of the set A.

It is worth noticing that, because the asymptotic factorization (1) holds for h tending to zero, from a practical point of view too large values of h may increase the estimation error and should be avoided. On the other hand, too small values of h should not be considered since, for small sample sizes, $\widehat{\phi}$ could be null by chance: indeed the ball $B(X_j, h)$ could not contain any sample points other than X_j. In other words, in practice h should be chosen in an interval $\mathcal{H} = [h_m, h_M]$ (see [2,3] for some data driven ideas on this issue).

Moreover, since (2) is a nonparametric estimator of ϕ, if the latter is specified in a parametric way, then it can be used to estimate its parameters. To do this, assume that ϕ is specified by some parametric relation being of the form $\phi \in \{\phi_\theta, \theta \in \Theta \subset \mathbb{R}^p\}$. Some examples of processes for which a parametric form of ϕ is available, are the finite dimensional processes and some infinite Gaussian ones (see the Introduction). In these cases, estimating θ furnishes a measure of the complexity extent of the underlying process. An estimate θ_n of θ is defined as the minimizer over a suitable compact subset Θ of $\mathbb{R}^p$ of a dissimilarity between the target ϕ_θ and the nonparametric estimator $\widehat{\phi}$ given by (2). In particular, as done in [3], one can consider the *centered cosine dissimilarity* between $g(\phi_\theta)$ and $g(\widehat{\phi})$ computed on the sample and defined by

$$\Delta(\widehat{\phi}, \phi_\theta) = 1 - \frac{\langle g(\phi_\theta), g(\widehat{\phi})\rangle^2}{\|g(\phi_\theta)\|^2 \|g(\widehat{\phi})\|^2}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the usual inner product and the associated norm respectively in $\mathcal{L}^2_{\mathcal{H}}$, and g is a suitable continuous real-valued function defined on $\mathbb{R}$. If ϕ_θ and $\widehat{\phi}$ are both bounded away from zero on $\mathcal{H}$, then ϕ_θ, $\widehat{\phi}$, $g(\phi_\theta)$ and $g(\widehat{\phi})$ are in $\mathcal{L}^2_{\mathcal{H}}$ and (3) is wellposed. From a practical point of view, g can be chosen as the identity function if the parametric family is the one of finite dimensional processes, and can be chosen as the logarithm for infinite dimensional Gaussian processes (see [2]). In conclusion, an estimator θ_n of θ is given by:

$$\theta_n = \arg\min_{\theta \in \Theta} \Delta(\widehat{\phi}, \phi_\theta).$$

3 Some Asymptotics

In this section asymptotic results on the estimator θ_n, namely the $\sqrt{n}$-consistency and the asymptotic normality, are collected. The detailed derivations can be found in [3] under a β-mixing type condition.

Before presenting the results, it is useful to introduce the main definitions to describe short time dependence of a sequence of random elements.

Definition 1. Let $(X_i)_{i\in\mathbb{N}}$ be a process taking values in $\mathcal{F}$, $\mathcal{M}_a^b$ be the σ-field generated by $\{X_i, a \leq i \leq b\}$ where $a \leq b \in \mathbb{N}$ and the β–mixing coefficients $(\beta_k)_{k\geq 0}$ be defined by

$$\beta_k = \sup_{a\in\mathbb{N}} \mathbb{E}\left[\sup_{A \in \mathcal{M}_{a+k}^{\infty}} \mid \mathbb{P}\left(A|\mathcal{M}_1^a\right) - \mathbb{P}(A) \mid \right].$$

The process $(X_i)_{i\in\mathbb{N}}$ is called β–mixing if $\beta_k \to 0$ as $k \to +\infty$.

Definition 2. The sequence $(X_i)_{i\in\mathbb{N}}$ is a 1-approximating functional on a process $(Z_i)_{i\in\mathbb{Z}}$ with constants $(a_\ell)_{\ell\geq 0}$ if $\mathbb{E}\left[d\left(X_1, \mathbb{E}\left[X_1|\mathcal{G}_{-\ell}^{\ell}\right]\right)\right] \leq a_\ell$, $\ell = 0, 1, 2, \ldots$ where $\lim_{\ell\to\infty} a_\ell = 0$ and $\mathcal{G}_{-\ell}^{\ell}$ is a σ-field generated by $Z_{-\ell}, \ldots, Z_\ell$.

Let denote θ_0 the true value of θ, and choose Θ such that $\theta_0 \in \Theta$ and suppose that one of the following two conditions hold:

1.a (X_i) is an i.i.d. sequence;
1.b (X_i) is a 1-approximating functional of a stationary, β-mixing sequence with constants $(a_l)_{l\geq 0}$, with $\beta_k = O(k^{-\beta})$ for $\beta \geq 8$ and $a_k = O(k^{-a})$ for $a = \max\{\beta + 3, 12\}$.

Moreover assume that:

2. X_i are distributed as X, a $\mathcal{F}$–valued random elements for which:
   i. $\varphi_\chi(h) > 0$, for any $h > 0$ and χ;
   ii. $\sup_\chi |\varphi_\chi(h)/\phi(h) - \psi(\chi)| \longrightarrow 0$, as $h \longrightarrow 0$;
   iii. $\phi(h)$ is continuous and increasing on $\mathcal{H}$.

Then, under assumptions 1. and 2., one gets the following results:

i. θ_n converges to θ_0 in probability, as $n \longrightarrow \infty$;
ii. if $\Delta(\theta) \in C^2(\Theta)$ and strictly convex over the set Θ, then $\sqrt{n}(\theta_n - \theta_0)$ is asymptotically distributed as a centred Gaussian random variable.

4 Confidence Interval for the Complexity Index

Thanks to the asymptotic normality of the estimator illustrated in the previous section, the confidence interval for θ at the level $1-\alpha$ takes form: $\theta_n \pm z_{1-\alpha/2}\ se\,(\theta_n)$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a $\mathcal{N}\,(0,1)$ and $se\,(\theta_n)$ is the standard error of θ_n.

Since an explicit form of the standard error of the estimator $se(\theta_n)$ is not available, it must be estimated by using resampling methods: in particular, a block–Jackknife approach can be adopted. The latter is based on Jackknife block-deleted point value $\theta_n^{(j)}$ evaluated by removing the block $B_j = (X_j, \ldots, X_{j+\ell-1})$, of a suitable length $1 \leq \ell \leq n$, from the dataset $(X_1, \ldots, X_n)$. The estimated variance is:

$$\widehat{Var}_{\mathrm{BJ}}(\theta_n) = \frac{\ell}{(n-\ell)N} \sum_{j=1}^{n-\ell+1} \left(\widetilde{\theta}_n^{(j)} - \theta_n\right)^2,$$

where $\widetilde{\theta}_n^{(j)} = (n\theta_n - (n-\ell)\theta_n^{(j)})/\ell$ is the j–th Jackknife pseudovalue of θ_n.

Acknowledgements. E.G. Bongiorno, L. Chan and A. Goia are members of Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. Bongiorno, E.G., Goia, A.: Some insights about the small ball probability factorization for Hilbert random elements. Statist. Sinica **27**(4), 1949–1965 (2017)
2. Bongiorno, E.G., Goia, A., Vieu, P.: Evaluating the complexity of some families of functional data. SORT **42**(1), 27–44 (2018)
3. Bongiorno, E.G., Chan, K.L., Goia, A.: Detecting the complexity of a functional time series. J. Nonparametric Stat. **36**(3), 600–622 (2024)
4. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Theory and practice. Springer, New York (2006)
5. Horvath, L., Kokoszka, P.: Inference For Functional Data With Applications. Springer, New York (2012)
6. Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis. CRC Press, Boca Raton, FL (2017)
7. Li, W.V., Shao, Q.M.: Gaussian processes: inequalities, small ball probabilities and applications. In: Stochastic Processes: Theory and Methods, pp. 533-597, vol. 19 of Handbook of Statist. North-Holland, Amsterdam (2001)
8. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, New York (2005)



Studying a Mixture of Functional Data Having Different Complexity

Enea G. Bongiorno, Lax Chan, and Aldo Goia(✉)

Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, Via Perrone 18, 28100 Novara, Italy
{enea.bongiorno,lax.chan,aldo.goia}@uniupo.it

Abstract. This paper presents the problem of distinguishing the components of a functional data process whose components have different complexities. A heuristic strategy is proposed whose performance is measured by simulations.

Keywords: Small-ball probability · complexity index · degrees of freedom

1 Introduction

Functional Statistics aims to develop and analyze methods to study functional data, that is data having infinite dimensional nature (such as curves, surfaces, images, etc.). For an overview, see [4–6] and [7]. In this framework, a recent topic is the study of the complexity of functional data. This notion, firstly introduced in [1], is related to the asymptotic factorization of the Small-Ball Probability (SmBP). In particular, given a random process X taking values in a semi-metric space (H, δ), assume that

$$\varphi(x,h) = \mathbb{P}(\delta(X,x) \leq h) \sim \psi(x)\,\phi(h) \qquad \text{as } h \to 0$$

where $\psi(x)$ is the so-called pseudo-density and $\phi(h)$ plays the role of a complexity function. In fact, it brings information on the complexity of X. As an instance, if X is a diffusion process, then $\phi(h) = C_1 h^{\gamma} \exp\left(-C_2/h^{\beta}\right)$ whereas if X is a finite d-dimensional process then $\phi(h) = c_d h^d$. The parameters γ, β, d are called complexity indexes of X, where d can be interpreted as the degree of freedom or the number of random sources generating the underlying probability law (see [3]). Usually, a common structure of complexity is assumed for all the observed elements in the dataset. This work considers processes defined as a weighted mixture of components having different number of random sources (and hence different complexities). This kind of mixture is a completely new concept in the functional framework and has not a counterpart in the multivariate context. After defining this new concept in Sect. 2, an algorithm to study the complexity mixture structure is presented in Sect. 3 and its performances are illustrated in Sect. 4 by means of simulated examples.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 243–247, 2025.
https://doi.org/10.1007/978-3-031-64431-3_41

2 Mixture of Processes with Different Complexity

Let H be a space of real functions defined on a compact set $\mathcal{T} \subset \mathbb{R}$ and δ a suitable semi-metric. Take a partition $\{\Omega_1, \Omega_2\}$ of the sample space Ω for which $0 < \pi = \mathbb{P}(\Omega_1) < 1$, and assume:

$$X(\omega, t) = X_1(\omega, t)\, \mathbb{I}_{\Omega_1}(\omega) + X_2(\omega, t)\, \mathbb{I}_{\Omega_2}(\omega), \qquad t \in \mathcal{T} \tag{1}$$

where the component X_g (with $g = 1, 2$) depends on d_g i.i.d. real random variables. Then, as $h \to 0$, the complexity function of each X_g has the form $\phi_g(h) \sim c_g h^{d_g}$ (with $c_g > 0$). In this setting, the SmBP of X is

$$\varphi(x, h) = \pi \varphi_1(x, h) + (1 - \pi)\, \varphi_2(x, h)$$

where $\varphi_g(x, h)$, $g = 1, 2$, is the g-th conditional SmBP. Assuming that $\varphi_g(x, h) \sim \psi_g(x)\, \phi_g(h)$ as $h \to 0$, the SmBP writes as a weighted sum of factors $\psi_g(x)\, \phi_g(h)$ and, up to an identifiability condition $\mathbb{E}[\psi_g(X)] = 1$, it can be proved that

$$\phi(h) \sim \pi c_1 h^{d_1} \qquad h \to 0 \tag{2}$$

where d_1 is the smallest complexity index in the mixture.

As an example, let $H = \mathcal{L}^2_{[0,1]}$, the Hilbert space of square integrable functions on $[0, 1]$ equipped with δ, the usual metric in H, and consider the process (1) whose components are defined by $X_g(\omega, t) = \sum_{j=1}^{d_g} \tau_{g,j}(\omega) \sqrt{\lambda_j} \xi_j(t)$, $t \in [0, 1], g = 1, 2$ where $\tau_{g,j}$ are continuous independent standardized real random variables, $\lambda_1 > \lambda_2 > \cdots > 0$ and $\{\xi_j\}$ is an orthonormal basis of H. Consequently, X_g is in $H_g = span(\xi_1, \ldots, \xi_g)$ with $\phi_g(h) = c_g h^{d_g}$ as $h \to 0$, whereas the complexity index of the leading term of X is d_1. In this example, the dimensionalities and the complexity indexes of the components coincide.

3 Evaluating the Complexity Mixture of a Sample

Let $\chi = (\chi_1, \ldots, \chi_n)$ be a sample drawn from X. The empirical estimators over χ of the SmBP $\varphi(x, h)$ and the complexity function $\phi(h)$ are:

$$\varphi_n(x, h) = n^{-1} \sum_{1 \le i < j \le n} \mathbb{I}_{\{\chi_i \in B(x,h)\}}$$

where $B(x, h)$ is a ball with center x and radius h (small enough) and

$$\phi_n(h) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \mathbb{I}_{\{\chi_i \in B(\chi_j,h)\}}. \tag{3}$$

Consider the cosine-dissimilarity measure $\Delta(\phi_n, \phi) = 1 - \langle \phi, \phi_n \rangle^2 / \left(\|\phi\|^2 \|\phi_n\|^2 \right)$ where $\phi(h) = c_d h^d$ with $d \in \mathbb{N}$. Due to (2), an estimator of the leading complexity index d_1 of X is the minimizer of Δ over is a suitable compact subset $D \subset \mathbb{R}$ that one assumes to contain the true value of d_1:

$$\widehat{d_1} = \arg\min_{d \in D} \Delta(\phi_n, \phi). \tag{4}$$

Theoretical properties of the nonparametric estimators (3) and (4) are derived in [2] in the i.i.d. case and in [3] under a β-mixing dependency structure. At this stage one may wonder if it is possible also to estimate d_2, the complexity index of the second group, and the mixture weight π. In principle the answer is affirmative if it is possible to separate X_1 from X_2. Note that, thanks to equation (1), it is possible to partition the dataset as $\chi = \chi_1 \cup \chi_2$, where χ_g is the set of observed curves χ_g drawn from the component X_g. Thus, if one is capable of identifying the elements of χ_1 then, by removing them from χ and computing the estimators (3) and (4) on the remaining part of the sample, one can obtain $\widehat{d}_2$, an estimate of the complexity index of the second component X_2. Moreover, an estimate of the mixture weight is given by $\widehat{\pi} = n_1/n$, where $n_1 = \#\chi_1$.

To identify the elements of χ_1 one can evoke and adapt the naive Bayes rule: taken $x = X(\omega)$, if $\omega \in \Omega_1$ then, it is possible to prove that there exists h_0 such that for any $h < h_0$ it holds

$$\mathbb{P}(\Omega_1 \mid \|X - x\| \leq h) \geq \mathbb{P}(\Omega_2 \mid \|X - x\| \leq h)$$

that is equivalent to

$$\pi\varphi_1(x, h) \geq (1 - \pi)\varphi_2(x, h). \tag{5}$$

Combining (5) with (2), one concludes that the global SmBP $\varphi(x, h)$ dominates $(1 - \pi)\varphi_2(x, h)$ for any $h < h_0$. This leads to the empirical rule: select h small enough to guarantee that the SmBP factorises and the indicator function in (3) does not vanish systematically, if $\varphi_n(\chi_i, h)$ is sufficiently far from zero, then χ_i is assigned to χ_1, otherwise to χ_2.

It emerges that the selection of a suitable bandwidth h is a crucial step for which a theoretical rule is not yet available. Since the volume must be strictly positive and then the only term able to discriminates groups is the pseudo-density, one can use as starting point the classical multivariate plug-in selector introduced in [8].

4 Simulation Study

This section is devoted to illustrate the performances of the proposed method by simulations. The study combines the following experimental settings:

- The process is a mixture of two components where the real random sources enter in the following nonlinear way (the complexity indexes used are $d_1 = 1$ and $d_2 = 3$):

$$X_1(t) = \tau_{1,1}t + \tau_{1,1}\cos(\tau_{1,1}2\pi t)$$
$$X_2(t) = \tau_{2,1}t + \tau_{2,2}\cos(\tau_{2,3}2\pi t)$$

  where $t \in [0, 1]$ and, for any g and j, $\tau_{g,j} \sim \mathcal{N}(4, 1)$ and independent;
- Sample size $n = 200, 400, 600$;
- Weight of the first component $\pi = 0.3, 0.5, 0.7$.

Table 1. Estimates of the complexity indexes d_1 and d_2, and the mixture weight π in a nonlinear case with $d_1 = 1$ and $d_2 = 3$.

π	n	$\widehat{d_1}$	$\widehat{d_2}$	$\widehat{\pi}$
0.3	200	1.343 (0.479)	2.957 (0.208)	0.355 (0.036)
	400	1.046 (0.210)	3.039 (0.199)	0.377 (0.027)
	600	1.002 (0.045)	3.001 (0.055)	0.310 (0.032)
0.5	200	1.002 (0.045)	2.838 (0.522)	0.483 (0.038)
	400	1.000 (0.000)	2.974 (0.165)	0.471 (0.017)
	600	1.000 (0.000)	2.986 (0.118)	0.485 (0.013)
0.7	200	1.000 (0.000)	2.597 (0.530)	0.618 (0.024)
	400	1.000 (0.000)	2.630 (0.483)	0.657 (0.014)
	600	1.000 (0.000)	2.875 (0.331)	0.672 (0.010)

For each scenario, 1000 Monte Carlo simulations are run and the mean and the standard deviation of $\widehat{d_1}$, $\widehat{d_2}$, $\widehat{\pi}$ are computed: the outcomes are collected in Table 1 (the standard deviations are in parenthesis).

Reading these results it emerges that the proposed methodology performs well in a variety of controlled settings. In particular, the estimates of the first dimension d_1 are very good varying both the mixture weight π and the sample size n. A slight overestimation appears when $\pi = 0.3$ and $n = 200$: this is related to the small number of curves in the first group. Also the estimates of the complexity index of the second group d_2 are generally good, with a variability that decreases as the sample size increases. An underestimate occurs when $\pi = 0.7$ and $n = 200$, due to a low availability of curves in the second group. Finally, the estimates of π produce good performances with an increase of precision as the sample size increases.

Acknowledgements. E.G. Bongiorno, L. Chan and A. Goia are members of Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. Bongiorno, E.G., Goia, A., Vieu, P.: Evaluating the complexity of some families of functional data. SORT **42**(1), 27–44 (2018)
2. Bongiorno, E.G., Goia, A., Vieu, P.: Estimating the complexity index of functional data: some asymptotics. Stat. Probab. Lett. **161**, 108731 (2020)
3. Bongiorno, E.G., Chan, K.L., Goia, A.: Detecting the complexity of a functional time series. J. Nonparametric Stat. **36**(3), 600–622 (2023)
4. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Theory and practice. Springer, New York (2006)
5. Horvath, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer, New York (2012)

6. Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis. CRC Press, Boca Raton, FL (2017)
7. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, New York (2005)
8. Wand, M.P., Jones, C.: Multivariate plug-in bandwidth selection. Comput. Stat. **9**(2), 97–116 (1994)



Partial M-Quantile Regression for Predictive Mantainance

Riccardo Borgoni[1(✉)], Enrico Fabrizi[2], Nicola Salvati[3], Francesco Schirripa Spagnolo[3], and Diego Zappa[4]

[1] Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano Bicocca, Milan, Italy
riccardo.borgoni@unimib.it
[2] Dipartimento di Scienze Economiche e Sociali, Università Cattolica del S. Cuore, Piacenza, Italy
[3] Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy
[4] Dipartimento di Scienze statistiche, Università Cattolica del S. Cuore, Milan, Italy

Abstract. In Industry 4.0 factories, innovative prediction tools are adopted so that data can be systematically processed into information that can explain uncertainties and support decisions. Predictive manufacturing systems begin with acquiring data from monitored assets using appropriate sensors to extract various signals. These signals can then be integrated with historical data into extensive datasets containing a multitude of variables. Consequently, addressing the challenge of reducing dimensionality becomes of paramount importance. Dimension reduction techniques such as partial least squares (PLS) have recently gained attention to deal with the problem of big datasets with a large number of correlated variables. Standard PLS approaches confine the estimation to examining only average effects, resulting in an insufficient portrayal. In this paper, we combine the standard PLS technique with M-quantile regression. The proposed approach aims at offering a more comprehensive view of the effect of various dimensions on the degradation of etching equipment in the microchip fabrication process.

Keywords: Partial Least Square · High Dimensional Data · Microelectronics

1 Introduction

The fourth industrial revolution also named Industry 4.0 was first mentioned in 2011 as a proposal for the development of a new concept of German economic policy based on high-tech strategies [7]. In Industry 4.0 factories, innovative prediction tools are adopted so that data can be systematically processed into information that can explain uncertainties and support decisions. Process engineers, instead of working with a physical model of the system which may be difficult to manage, prefer model-free approaches and use advanced techniques to monitor, control, and optimize the performance of the industrial process,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 248–253, 2025.
https://doi.org/10.1007/978-3-031-64431-3_42

essentially based on huge amounts of measurements. Predictive manufacturing system starts with data acquisition of the monitored assets using appropriate sensors to extract various signals such as temperature, vibration or pressure. According to [6] industrial and process plants traditionally employ two types of maintenance management: run-to-failure or preventive maintenance. In run-to-failure maintenance, action for repairing equipment is performed only when the equipment has broken down or been run to the point of failure and no attempt is made to anticipate maintenance requirements hence, a plant must be able to react to all possible failures. This forces the maintenance department to maintain extensive spare parts in the plant.

Preventive maintenance is based on hours of operation. All preventive maintenance management programs assume that machines will degrade within a time frame typical of their typology and interventions a made on a scheduled basis. Modern predictive maintenance has a different philosophy. Predictive maintenance is a condition-driven preventive maintenance program that uses possibly huge amount of data for monitoring the system to evaluate its condition and efficiency. Machine learning and statistical learning techniques are nowadays the main tools by which predictive maintenance operates in practice. The large majority of these approaches, however, aim at assessing the impact of the potentially influential factors on average, that is focusing on the expected value of the outcome variable or variables. Indeed in many circumstances, the tail of the distribution might be more interesting reflecting the critical status of the equipment.

In this paper, we consider a predictive maintenance problem in the microelectronics industry. In microelectronics, semiconductor devices are developed by highly integrated sequences of several technological steps on circular-shaped working substrates called wafers. In the final production phase, wafers are cut into small items, called dice, and then packaged into chipsets. Hereinafter, we consider the dry-etching phase of the integrated circuit production process. One of the preliminary steps of chipset fabrication process consists of the deposition of a thin SiO_2 film over the wafers. Etching technologies aim to transfer the patterns defined by photolithographic steps to the surface of the wafers by selective material removal. The term *Plasma Etching* refers to the process by which a plasma reactor is used to remove excess material from the surface of semiconductors allowing precise control of the profile dimensions and structures.

The type of plasma reactors considered in this paper is characterized by the electrical excitation supplied to the incoming gas. Indeed, they are powered by an electrical signal called radio frequency. Radio frequency discharges aim to create an electromagnetic field inside the plasma chamber and consequently induce the ionization of the gas. We consider quartz degradation monitored during the etching phase of chip fabrication. Quartz degradation is measured in terms of the reflected power (RF) i.e. the portion of radio frequency power that is reflected back from the plasma inside the etching chamber towards the power source. High levels of RF can indicate poor plasma coupling and can lead to inefficient etching, non-uniformity, or damage to the equipment. Therefore,

monitoring and controlling reflected power is essential for achieving high-quality and reliable etching processes

In this work, we generalize the Partial Quantile Regression (PQR) to the M-quantile setting, and apply this approach to assess the differential effects of a large number of covariates on the RF distribution aiming at identifying those parameters that mainly impact on the highest levels (M-quantiles) of RF. To account for potentially high correlations among predictors, we combine M-quantile regression with a partial least squares approach.

2 The Partial M-Quantile Model

Partial Least Squares (PLS) [3,8] is a dimension reduction method for deriving a low-dimensional set of features from a large set of variables. It can be used in situations where the variables in data are highly correlated and/or in the case of high dimensional data. Moreover, PLS could be considered as a *supervised* alternative to Principal Component Analysis because it uses the response to construct the directions of the new dimensions [4]. Indeed, the PLS scores are calculated to maximize the covariance between the scores of an independent block and the scores of a dependent block. However, when the interest is in extreme events, as for the aim of our work, the standard PLS approach cannot be used because it just allows us to find an average relationship between the response and predictors. In the classical regression framework, an alternative to conditional mean regression is Quantile Regression (QR) proposed by [5] which allows evaluating the impacts of predictors on the outcome at different quantile levels. [2] combined PLS with QR developing the Partial Quantile Regression (PQR). This methodology parallels the procedure for PLS by maximising a quantile covariance rather than the usual covariance which is appropriate for predicting a mean value. M-quantile regression [1] generalizes both quantile and expectile regression using the M-estimation approach. In particular, depending on the choice of the loss function, M-quantiles may be reduced to ordinary quantiles and expectiles.

2.1 The M-Quantile Regression Model

M-quantile regression has been introduced by [1]. It extends the concepts of M-estimation to model the relationship between the dependent variable and its predictors at various parts of the conditional distribution.

Formally, the M-quantile of order q for the conditional density of Y given the set of covariates X is defined as the solution $MQ_y(q|X,\psi)$ of the integral equation $\int \psi_q[y - MQ_y(q|X,\psi)]f(y|X)dy = 0$ where $\psi_q(\epsilon_q) = 2\psi(\epsilon_q/\sigma_q)\{(1-q)I(\epsilon_q \leq 0) + qI(\epsilon_q > 0)\}$ and σ_q is a scale parameter.

Let $\mathbf{X}$ be the data matrix of size $n \times p$ of values of X and let y the n-vector of Y values. Given $\mathbf{X}_i$, i.e. the value of X for unit i, $i = 1, \ldots, n$, the linear M-quantile regression model is defined by

$$MQ_{y_i}(q|\mathbf{X}_i,\psi) = \mathbf{X}_i^T \beta_{\psi q}, \tag{1}$$

where $\beta_{\psi q}$ represents a vector of unknown parameters at the qth M-quantile obtained by using iterative weighted least square algorithm and ψ is Huber proposal 2 influence function:

$$\psi(u) = \begin{cases} u & \text{if } -c \leq u \leq c \\ c\,\text{sign}(u) & \text{otherwise,} \end{cases} \tag{2}$$

where c is a tuning constant bounded away from zero. By setting the value of c, it is possible to trade robustness for efficiency in the M-quantile regression (MQ) model.

2.2 The PMQ Algorithm

PLS can be considered as a covariance optimization problem. Therefore, the preliminary step is to define a sort of M-quantile covariance metric linked to the M-quantiles to be optimized, instead of the traditional covariance, that is linked to the mean. For this reason, we generalize the quantile covariance defined by [2], defining the so-called M-quantile covariance:

$$Cov_{\psi;\tau}(\mathbf{X}, Y) = var(\mathbf{X})\beta_{(\psi;\tau)} \tag{3}$$

where $\beta_{(\psi;\tau)}$ is the regression coefficient obtained by fitting a standard MQ regression at a given τ with Y as the dependent variable and the variables in $\mathbf{X}$ as regressors; $var(\mathbf{X})$ is the variance of the predictors. This is a covariance-type measure of the association between Y and $\mathbf{X}$ formulated in terms of predicting the M-quantile τ of Y from $\mathbf{X}$.

Moreover, we slightly modified the PLS (and PQR) algorithm by calculating the y-loadings by using an M-quantile regression instead of the standard OLS formulation. This allows us to have robust PLS regression coefficient estimates. The proposed algorithm for the PMQ is as follows:

Start by setting $\mathbf{X}_0 = \mathbf{X}$ and $y_0 = y$.

At a given $\tau \in (0, 1)$, repeat for $h = 1, \ldots, H$

1. Compute $s_{\tau,h} = Cov_{\psi;\tau}(\mathbf{X}_{h-1}, y_{h-1})$, the sample covariance, and normalize it: $w_{\tau,h} = s_{\tau,h}/||s_{\tau,h}||$ (this is equivalent to obtain the eigen decomposition of $s_{\tau,h}$ and take $w_{\tau,h}$ as the eigenvector associated to the largest eigenvalue)
2. Compute the PLS component: $t_{\tau,h} = \mathbf{X}_{h-1}w_{\tau,h}$
3. Calculate the X-loading vector as: $p_{\tau;h} = \mathbf{X}_{h-1}^T t_{\tau,h}/t_{\tau,h}^T t_{\tau,h}$
4. Calculate the y-loading: $b_{\tau;h} = y_{h-1}^T \delta_\tau t_h / t_h^T \delta_\tau t_h$, where δ_τ is a vector of weights obtained from the iterative reweighted least squares algorithm used to fit the M-quantile regression at τ.
5. Deflate the predictors $\mathbf{X}$ and the response variables as: $\mathbf{X}_h = \mathbf{X}_{h-1} - t_{\tau,h}p_{\tau,h}$ and $y_h = y_{h-1} - b_{\tau,h}t_{\tau,h}$

At the end, the prediction model for y at the quantile of interest τ can be written as:

$$\hat{y} = \mathbf{T}_\tau \mathbf{b}_\tau,$$

Recall that at each τ the PLS components can also be expressed in terms of the original variables: $\mathbf{T}_{\tau,h} = \mathbf{X}\mathbf{W}^{\star}_{\tau,h}$, where $\mathbf{W}^{\star}_{\tau,h} = \mathbf{W}_{\tau,h}(\mathbf{P}^{T}_{\tau,h}\mathbf{W}_{\tau,h})^{-1}$.

Therefore, the y-loading of the PLS regression can be expressed in terms of the original predictors as:

$$\mathbf{b}^{\star}_{\tau} = \mathbf{X}\mathbf{W}^{\star}_{\tau}\mathbf{b}_{\tau} \tag{4}$$

To select the number of components at each quantile of interest, we use k-fold Cross-Validation, with $k = 7$ and the Root Mean Squared Error of Prediction (RMSEP) for validation.

3 The PMQ for Predictive Maintenance: Results

The methodology presented above has been applied to data taken from the dry-etching phase of the integrated circuit production process.

Results are reported in Table 1. The last two columns show the estimated coefficients when the PLS approach is not used, and the M-quantile regression is directly estimated on the original variables. More specifically, M-quantiles of order 0.5 and 0.9 were considered in the table. As expected, the high correlation between certain pairs of variables in the dataset has resulted in unstable estimates and unreliable values have been obtained. Utilizing the PLS approach allows one to circumvent this issue, as demonstrated in columns two and four, where the regression coefficients have been reconstructed after applying the PMQ approach. To estimate the standard error of the coefficients we employ a bootstrap approach where the entire algorithm has been applied to each bootstrap replicate of the sample. This takes into account the variability that arises in the computation of the scores (step 1. and 2. in the PMQ algorithm) in addition to the variability of the estimates of the regression coefficients. With this approach, the p-values reported in columns 3 and 5 are more accurately estimated.

Table 1. Estimates of the regression coefficients and their Bootstrap p-values.

	PMQ				MQ	
Variable	$\mathbf{b}^{\star}_{0.5}$	P-value$_{0.5}$	$\mathbf{b}^{\star}_{0.9}$	P-value$_{0.9}$	$\widehat{\beta}_{0.5}$	$\widehat{\beta}_{0.9}$
A1	0.0377	0.203	0.0439	0.091	0.0242	0.0234
A2	–0.0098	0.735	–0.0414	0.079	–0.0796	–0.0742
A3	0.0158	0.158	0.0310	0.015	0.0188	0.0243
B1	0.0538	0.057	0.0262	0.320	0.0237	0.0416
B2	–0.0465	0.078	–0.0355	0.144	–0.0770	–0.0968
B3	0.0445	0.030	0.0299	0.116	0.0361	0.0495
B4	–0.0419	0.001	–0.0208	0.141	–0.0360	–0.0426
C1	0.0460	0.064	0.0396	0.074	0.0982	0.0759
C2	0.5162	0.000	0.5371	0.000	2285.1	2166.7
C3	–0.0342	0.012	–0.0273	0.065	–0.0306	–0.0362
C4	–0.0221	0.035	–0.0177	0.094	–0.0224	–0.0154
C5	0.0316	0.087	0.0325	0.040	0.1426	0.1463

In Table 1 the estimated coefficients of 12 out of the 35 variables considered in the analysis are considered. It can be observed that not only do the sizes of the coefficients change when moving from M-quantile 0.5 to M-quantile 0.9 (increasing in modulus, roughly, from 15% to 320%), but also the associated p-values differ. Specifically, the variables labelled as A in the table exhibit significance at higher M-quantiles, indicating potential parameters engineers should monitor and act upon to control higher values of reflected power, highlighting more critical situations. The variables labelled as C in the table exhibit significance at the 0.9 M-quantile and remain significant even at lower M-quantiles. Their impact extends across the entire conditional distribution of reflected power, warranting caution from process engineers. In particular, variable C2 represents the time since the process started. This dimension impacts at different levels of the response distribution in a similar manner as it might be expected. Alternatively, variable A2 represents the impact of a secondary generator in the etching chamber that tends to compensate for an excess of RF. This variable has a more noticeable effect on the right tail of the RF distribution (the estimated coefficient at the 0.9 M-quantile being about 300% higher in modulus than the coefficient at the M-median). This suggests that setting this variable may help in preventing high values of RF. The remaining variables were not found to be significant at either the 0.5 or the 0.9 M-quantile and their estimated coefficients were not reported in Table 1.

References

1. Breckling, J., Chambers, R.: M-quantiles. Biometrika **75**, 761–771 (1988)
2. Dodge, Y., Whittaker, J.: Partial quantile regression. Metrika **70**, 35–57 (2009)
3. Helland, I.S.: On the structure of partial least squares regression. Commun. Stat. Simulation Comput. **17**, 581–607 (1988)
4. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning. Springer (2013)
5. Koenker, R., Bassett, G., Jr.: Regression quantiles. Econometrica **46**, 33–50 (1978)
6. Mobley, R.K., An introduction to predictive maintenance. Second Edition. Butterworth-Heinemann (2002)
7. Mosconi, F.: The new European Industrial Policy: Global competitiveness and the Manufacturing Renaissance. Routledge, London, England (2015)
8. Wold, H.: Estimation of principle components and related models by iterative least squares. In: Keishnaiah, P.R., (eds.) Multivariate Analysis. Academic Press (1966)
9. Zonta,T, André da Costa, T., da Rosa Righi, R., de Lima, M.J., Silveira da Trindade E., Li, GP.: Predictive maintenance in the industry 4.0: a systematic literature review. Comput. Indus. Eng. **150**, 106889 (2020)



Completion of Partially Observed Multivariate Functional Data

Marco Borriero[1(✉)], Luigi Augugliaro[2], Salvatore Latora[2], and Veronica Vinciotti[3]

[1] Department of Statistics, Computer Science and Applications, University of Florence, Florence, Italy
marco.borriero@unifi.it
[2] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
{luigi.augugliaro,salvatore.latora}@unipa.it
[3] Department of Mathematics, University of Trento, Trento, Italy
veronica.vinciotti@unitn.it

Abstract. Functional data often present missing values. These pose challenges when performing functional principal components analysis, since it is not possible to compute the scores from the data. The existing method by Kraus (2015) allows to perform functional data completion in a univariate setting. In this paper, we propose an extension to the multivariate setting and show how statistical dependences between the functional curves aid in the imputation of missing data.

Keywords: Multivariate functional data · Missing data · Partial separability

1 Introduction

In the last years, the study of multivariate data in the form of functions or curves, referred to as multivariate functional data, has become popular. Formally, multivariate functional data are characterized by the fact that every observation contains measurements relative to a multivariate random variable, as time varies. The problem of missing data occurs frequently in the experiments in which functional data are recorded and this gives rise to functional data that are partially observed. In these cases, it is necessary to develop procedures for the completion of the missing parts of the curves.

An important tool is the Karhunen-Loève expansion ([3]), which permits to express functional data as a linear combination of a given basis where the coefficients, which are called *functional principal component scores*, contain all the information of the data. The main difficulty that arises when data are not completely observed is that these scores cannot be computed from data. To address this challenge, [1] proposes a statistical approach in which the missing parts of the scores of univariate functional data are imputed via a functional

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 254–259, 2025.
https://doi.org/10.1007/978-3-031-64431-3_43

linear regression with the observed data as predictors. In our work, we extend this procedure to the multivariate setting, so as to exploit the intrinsic correlations between the functional curves.

2 Multivariate Functional Data Under Partial Separability

In the multivariate setting, the functional data is composed of n observations drawn from a zero-mean, measurable and square integrable process $\{\mathbf{X}(t) : t \in [0,1]\}$, where $\mathbf{X}(t) \in \mathbb{R}^p$. It is assumed that $E[X_j(t)] = 0$, for each $t \in [0,1]$. The process $\mathbf{X}$ lies in the space $L_2^p([0,1])$ equipped with inner product $\langle \mathbf{f}, \mathbf{g} \rangle_p = \sum_{i=1}^p \langle f_i, g_i \rangle$, for $\mathbf{f}, \mathbf{g} \in L_2^p([0,1])$.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent and identically distributed realizations of the process. These constitute the functional data. A key tool is the covariance operator. Indeed, under a Gaussianity assumption, this determines completely the distribution of the data. The covariance operator is a function $\mathcal{G} : L_2^p([0,1]) \to L_2^p([0,1])$ which represents an extension of the covariance matrix of a multivariate process. It can be thought as a matrix of operators $\{\mathcal{G}_{jl}\}_{j,l=1}^p$, in which each entry $\mathcal{G}_{jl} : L_2([0,1]) \to L_2([0,1])$ is an integral operator defined as $\mathcal{G}_{jl}(f) := \int_0^1 g_{jl}(\cdot, t) f(t) dt$ for any $f \in L_2([0,1])$, where $g_{jl}(s,t) = Cov\left(X_j(s), X_l(t)\right)$ is the kernel function. In [2], an assumption on the covariance operator, called *partial separability*, is proposed. Formally, a covariance operator $\mathcal{G} : L_2^p([0,1]) \to L_2^p([0,1])$ is said to be partially separable if there are an orthonormal basis $\{e_{kj}\}_{j=1}^p$ of $\mathbb{R}^p$ and an orthonormal basis $\{\varphi_k\}_{k=1}^{+\infty}$ of $L_2([0,1])$ such that its eigenfunctions are of the form $\{e_{kj}\varphi_k\}_{k,j}$.

[2] prove that $\mathcal{G}$ is partially separable if and only if $\mathcal{G} = \sum_{k=1}^{+\infty} \Sigma_k \varphi_k \otimes \varphi_k$, where Σ_k are $p \times p$ symmetric and positive definite matrices. Moreover, the assumption of partial separability implies the following multivariate Karhunen-Loève expansion: $\mathbf{X} = \sum_{k=1}^{+\infty} \boldsymbol{\xi}_k \varphi_k$, where $\boldsymbol{\xi}_k = (\langle X_1, \varphi_k \rangle, \ldots, \langle X_p, \varphi_k \rangle)^T$, $Var(\boldsymbol{\xi}_k) = \Sigma_k$ and the multivariate scores are mutually uncorrelated. We assume also that the process is Gaussian which implies that $\boldsymbol{\xi}_k$ follows a multivariate Gaussian distribution.

3 Missing Data Imputation for Multivariate Functional Data

The problem that arises with partially observed functional data is that the scores cannot be computed directly from the data. In particular, for each k, the jth score is defined as a sum of an observed and missing part:

$$\xi_{jk} = \int_{O_j} X_j(t)\varphi_k(t)dt + \int_{M_j} X_j(t)\varphi_k(t)dt = \langle X_j, \varphi_k \rangle_{O_j} + \langle X_j, \varphi_k \rangle_{M_j} = \xi_{jk}^{O_j} + \xi_{jk}^{M_j},$$

where O_j and M_j are disjoint subintervals of $[0,1]$ where $X_j(t)$ is observed or missing, respectively. Under the assumption that data are missing completely at

random, the method proposed in [1] relies on the idea of imputing the missing part of the score, i.e., $\xi_{jk}^{M_j}$, using the best linear predictor of the conditional expectation $E[\xi_{jk}^{M_j}|\mathbf{X}^O]$. This assumes the form $\langle \beta_{jk}, X_j \rangle_{O_j}$, for all $j = 1, \ldots, p$, and $\beta_{jk} \in L_2(O_j)$ is chosen so that

$$\hat{\beta}_{jk} = arg \min_{\beta_{jk} \in L_2(O_j)} \frac{1}{2} E\left[\left(\xi_{jk}^{M_j} - \langle \beta_{jk}, X_j \rangle_{O_j} \right)^2 \middle| X_j^{O_j} \right].$$

In [1] it is proved that $\hat{\beta}_{jk}$ of the above problem satisfies also

$$\mathcal{G}_{jj}^{O_j O_j} \hat{\beta}_{jk} = \mathcal{G}_{jj}^{O_j M_j} \varphi_k^{M_j}, \tag{1}$$

where $\mathcal{G}_{jj}^{O_j O_j} : L^2(O_j) \to L^2(O_j)$ is the covariance operator of the curve $X_j^{O_j}$, $\mathcal{G}_{jj}^{O_j M_j} : L^2(M_j) \to L^2(O_j)$ is the covariance between $X_j^{M_j}$ and $X_j^{O_j}$ and $\varphi_k^{M_j}$ is the restriction of the basis φ_k on M_j.

The limitation of this approach is that the p components of the score $\boldsymbol{\xi}_k$ are treated as univariate random variables. As such, statistical dependences between the different curves are not exploited in the imputation of the missing values. Put it differently, the imputation does not exploit the complete information carried by the data. In contrast to this, the solution that we propose, exploits the correlations of the different variables in order to gain more information from the data, leading to an expected improvement in the imputations of the scores.

Going into the details, under the assumption of partial separability, the scores $\boldsymbol{\xi}_k$ follow a multivariate Gaussian distribution with zero mean and covariance matrix Σ_k. However, the distribution of the scores conditional on $\mathbf{X}^O$ is not known. Therefore, we assume that $\boldsymbol{\xi}_k|\mathbf{X}^O \sim N_p(\boldsymbol{\mu}_k^{M|O}, \Sigma_k^{M|O})$ with mean $\boldsymbol{\mu}_k^{M|O} = E[\boldsymbol{\xi}_k|\mathbf{X}^O]$ and precision matrix $\Theta_k^{M|O} = (\Sigma_k^{M|O})^{-1}$. The density function of $\boldsymbol{\xi}_k|\mathbf{X}^O$ is then given by

$$f(\boldsymbol{\xi}_k|\mathbf{X}^O) = (2\pi)^{\frac{p}{2}} \left(\det \Theta_k^{M|O} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\boldsymbol{\xi}_k - \boldsymbol{\mu}_k^{M|O} \right)^T \Theta_k^{M|O} \left(\boldsymbol{\xi}_k - \boldsymbol{\mu}_k^{M|O} \right) \right),$$

for each k and the profile log-likelihood function of the mean $\boldsymbol{\mu}_k^{M|O}$ is

$$l(\boldsymbol{\mu}_k^{M|O}|\Theta_k^{M|O}) \propto -\frac{1}{2} \left(\boldsymbol{\xi}_k^O + \boldsymbol{\xi}_k^M - \boldsymbol{\mu}_k^{M|O} \right)^T \Theta_k^{M|O} \left(\boldsymbol{\xi}_k^O + \boldsymbol{\xi}_k^M - \boldsymbol{\mu}_k^{M|O} \right).$$

Since the mean $\boldsymbol{\mu}_k^{M|O}$ can be expressed as the sum of the expected values of the missing and of the observed parts of the scores conditioning on the observed data, it follows that

$$\boldsymbol{\mu}_k^{M|O} = \boldsymbol{E}[\boldsymbol{\xi}_k|\mathbf{X}^O] = \boldsymbol{E}[(\boldsymbol{\xi}_k^O + \boldsymbol{\xi}_k^M)|\mathbf{X}^O] = \boldsymbol{\xi}_k^O + \boldsymbol{E}[\boldsymbol{\xi}_k^M|\mathbf{X}^O] = \boldsymbol{\xi}_k^O + \boldsymbol{\mu}_k^M$$

and so the profile log-likelihood function of the mean of the missing term $\boldsymbol{\xi}_k^M$ evaluated at $\boldsymbol{\mu}_k^{M|O}$ is given by

$$l(\boldsymbol{\mu}_k^{M|O}|\Theta_k^{M|O}) \propto -\frac{1}{2} (\boldsymbol{\xi}_k^M - \boldsymbol{\mu}_k^M)^T \Theta_k^{M|O} (\boldsymbol{\xi}_k^M - \boldsymbol{\mu}_k^M). \tag{2}$$

Let $\hat{\boldsymbol{\xi}}_k^M$ be the maximum likelihood estimator of $\boldsymbol{\xi}_k$. This cannot be computed directly since the profile log-likelihood (2) cannot be computed in closed form due to the missing parts of the scores being unknown. Thus, we define the objective function as the expectation of the negative log-likelihood function conditional to the observed data $\mathbf{X}^O$ and given the current estimates $\{\hat{\Theta}_k\}_k$. This can be written as

$$\begin{aligned}\mathcal{F}(\boldsymbol{\mu}_k^M) &= E[-l(\boldsymbol{\mu}_k^{M|O}|\Theta_k^{M|O})|\mathbf{X}^O;\{\hat{\Theta}_k\}_k] \\ &\propto \frac{1}{2} tr\left(\hat{\Theta}_k E\left[(\boldsymbol{\xi}_k^M - \boldsymbol{\mu}_k^M)(\boldsymbol{\xi}_k^M - \boldsymbol{\mu}_k^M)^T \middle| \mathbf{X}^O\right]\right),\end{aligned}$$

and the aim is to find $\hat{\boldsymbol{\xi}}_k^M \in \boldsymbol{R}^p$ that minimizes it. More in detail, we define the imputation of multivariate scores by generalizing the idea proposed by [1], in which the conditional expectation $\boldsymbol{\mu}_k^M$ is estimated by $\hat{\boldsymbol{\mu}}_k^M = \sum_{l=1}^p \langle \hat{\beta}_{jk,l}, X_l \rangle_{O_l}$ for all $j = 1, \ldots, p$ where $\hat{\boldsymbol{\beta}}_{jk} = (\hat{\beta}_{jk,1}, \ldots, \hat{\beta}_{jk,p})^T \in L_2^p(O)$ is such that:

$$\hat{\boldsymbol{\beta}}_{jk} = arg \min_{\boldsymbol{\beta}_{jk} \in L_2^p(O)} \frac{1}{2} E\left[\left(\xi_{jk}^{M_j} - \sum_{l=1}^p \langle \beta_{jk,l}, X_l^{O_l} \rangle_{O_l}\right)^2 \middle| \mathbf{X}\right].$$

It is possible to prove that $\hat{\boldsymbol{\beta}}_{jk}$ minimizes $\mathcal{F}(\boldsymbol{\mu}_k^M)$ and this is in analogy to (1) in the multivariate setting.

Thanks to the imputation of the missing terms of the scores, we can define our proposed EM type algorithm. In this algorithm, given imputed scores at a certain step of the iteration, we compute the next estimation of the covariance operator and use this for the imputations in the following iteration. In particular, given the current estimates of the basis system $\{\hat{\varphi}_k^{(h)}\}_k$, the E-step of the proposed algorithm imputes the missing terms of the multivariate scores $\{\hat{\boldsymbol{\xi}}_{ik}\}_k$, for all $i = 1, \ldots, n$. After that, it computes the sample covariance matrices of the scores, $S_k = \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\xi}}_{ik} \hat{\boldsymbol{\xi}}_{ik}^T$ and from them it computes the Q-function $\sum_k \log(\det(\Theta_k)) - tr(S_k \Theta_k)$.The M-step computes the next estimates of the precision matrices $\{\hat{\Theta}_k^{(h+1)}\}_k$ by maximizing the Q-function

$$\{\hat{\Theta}_k^{(h+1)}\}_k = arg \max_{\{\Theta_k\}_k} \sum_k \log(\det \Theta_k) - tr\left(S_k \Theta_k\right),$$

following by the calculation of the estimated covariance matrices $\{\hat{\Sigma}_k^{(h+1)}\}_k$ as the inverses of $\left\{\left(\hat{\Theta}_k^{(h+1)}\right)\right\}_k$. Finally, we define

$$\hat{X}_{ij}(t) = \begin{cases} X_{ij}(t) & if\ t\ \in O_{ij}, \\ \sum_{k=1}^K \hat{\xi}_{ijk} \hat{\varphi}^{(h)}(t) & if\ t\ \in M_{ij}. \end{cases}$$

and $\mathcal{F}_{jl}^{(h+1)} = \frac{1}{n} \sum_{i=1}^n \left(\hat{X}_{ij} - \hat{\mu}_j\right) \otimes \left(\hat{X}_{il} - \hat{\mu}_l\right)$. From this, we obtain the next basis system $\{\hat{\varphi}_k^{(h+1)}\}_k$ that will be used in the E-step of the next iteration

as the eigenfunctions of $\mathcal{F}^{(h+1)}$. Finally, thanks to the hypothesis of partial separability and using the covariance matrices $\{\hat{\Sigma}_k^{(h+1)}\}_k$, the next estimation of the covariance operator is given by $\hat{\mathcal{G}}^{(h+1)} = \sum_k \hat{\Sigma}_k^{(h+1)} \hat{\varphi}_k^{(h+1)} \otimes \hat{\varphi}_k^{(h+1)}$.

3.1 Simulation Study

We perform a simulation study to compare the proposed method with the univariate method of [1] in terms of accuracy in reconstructing a multivariate set of curves. The simulation setting is as follows. We fix the sample size $n = 50$ and the number of curves $p = 5$. Each curve is generated over a grid of $d = 50$ time points and we let $M_{ij} = \{t_k, \ldots, t_{k+w-1}\}$ with window $w = 20$ and index k sampled from a discrete uniform distribution. Moreover, we assume that 85% of curves are partially observed. To generate the curves we use a system of 3 Fourier basis. The true scores are generated from a multivariate Gaussian distribution with zero mean and covariance matrix generated under partial separability. In particular, we generate $p \times p$ precision matrices $\Theta_1, \ldots, \Theta_K$. Each precision matrix was generated as

$$\Theta_{ij} = \begin{cases} \rho^{|i-j|} & if \mid i-j \mid \leq id.diag \\ 0 & otherwise \end{cases},$$

for a correlation coefficient $\rho \in [0,1]$ selected as $\hat{\rho} = \arg\min_{\rho \in [0,1]} (det(\Theta) - det.min)^2$. In particular, *id.diag* determines the structure of the graph, and is set to 1, while *det.min* determines the strength of the dependence among the curves, and is set to 0.5.

Figure 1 shows the results of our study for 50 simulations. In particular, the mean squared error (MSE) was computed between the predicted scores and the true scores, for each curve and each simulation. Although the distribution of the error shows greater variability for our proposed method, in 50% of the cases

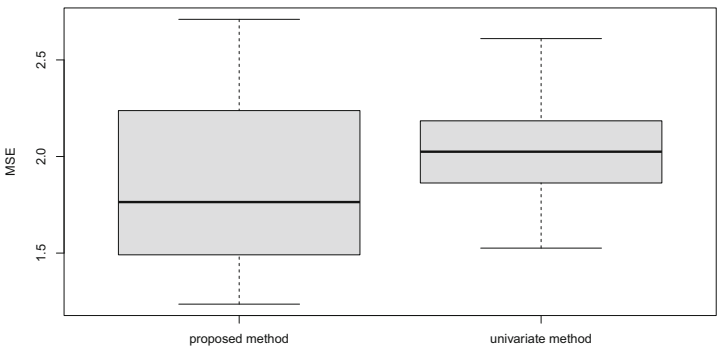


Fig. 1. Simulation study: MSE across 50 simulations between the predicted and true scores using the proposed multivariate approach (left) and the existing univariate approach of [1] (right).

the error is lower than that obtained using the univariate method. In practice, $\boldsymbol{\beta}_{jk}$ are found based on a ridge penalty parameter α. The choice of this tuning parameter was found to be responsible for the greater variability of the proposed method and will be investigated in the future.

4 Conclusions

In recent years, interest in functional data has increased considerably. Our work aims to provide a methodology for the reconstruction of partially observed multivariate curves. Preliminary results show that our proposed method performs well in the imputation of missing scores. We do not expect big changes to the results if the missing data are assumed to appear sparsely over the time domain because that is a simpler case of the one considered here. Finally, we expect that as n and d increase, the increased amount of information will lead to better estimates. In contrast to this, as p increases, there are more variables in the multivariate process to aid the functional completion of the curves, but at the expense of a greater computational cost.

References

1. Kraus, D.: Components and completion of partially observed functional data. J. Royal Stat. Soc. Series B (Stat. Methodol.) **77**(4), 777–801 (2015). https://doi.org/10.1111/rssb.12087.
2. Zapata, J., Oh, S.Y., Petersen, A.: Partial separability and functional graphical models for multivariate gaussian processes. Biometrika **109**(3), 665–681 (2021). https://doi.org/10.1093/biomet/asab046.
3. Ramsay, J.O. Silverman, B.W.: Functional Data Analysis, 2Nd Edn. Springer (2005)

Approximate Inference for the Bayesian Discrepancy Measure for Precise Statistical Hypotheses

Elena Bortolato[1], Francesco Bertolino[2], Monica Musio[2], and Laura Ventura[1(✉)]

[1] Department of Statistical Sciences, University of Padova, Padua, Italy
elena.bortolato@unipd.it, ventura@stat.unipd.it
[2] University of Cagliari, Cagliari, Italy
{bertolin,mmusio}@unica.it

Abstract. The aim of this paper is to introduce higher-order expansions for the Bayesian Discrepancy Measure for testing precise statistical hypotheses on a scalar parameter of interest. Connections with frequentist inference are highlighted when using objective matching priors. Moreover, an extension to the multivariate framework is mentioned.

Keywords: Bayesian Discrepancy Measure · credible set · evidence · higher-order asymptotics · matching priors · modified likelihood root · nuisance parameter · tail area probability

1 The Bayesian Discrepancy Measure

Consider a random sample $y = (y_1, \ldots, y_n)$ of size n from a random variable Y with parametric model $f(y;\theta)$, indexed by a parameter θ, with $\theta \in \Theta \subseteq \mathbb{R}^d$. Given a prior $\pi(\theta)$ on θ, Bayesian inference for θ is based on the posterior density $\pi(\theta|y) \propto \pi(\theta)L(\theta)$, where $L(\theta)$ represents the likelihood function based on $f(y;\theta)$. Of particular interest is the situation in which θ is partitioned as $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of interest and λ represents the remaining $(d-1)$ nuisance parameters. In such case, Bayesian inference for ψ is based on the marginal posterior density

$$\pi_m(\psi|y) = \int \pi(\psi, \lambda|y)\, d\lambda \propto \int \pi(\psi, \lambda) L(\psi, \lambda)\, d\lambda. \tag{1}$$

Here it is of interest to test the precise (or sharp) null hypothesis

$$H_0 : \psi = \psi_0 \quad \text{against} \quad H_1 : \psi \neq \psi_0. \tag{2}$$

In Bayesian hypothesis testing, the usual approach relies on the Bayes factor (BF), which measures the ratio of posterior to prior odds in favor of the null hypothesis, H_0. A high BF, or the weight of evidence $W = \log(BF)$, typically

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 260–265, 2025.
https://doi.org/10.1007/978-3-031-64431-3_44

leads us to favor H_0. However, using improper priors can render the BF undetermined, and in cases of precise null hypotheses, it can lead to the Jeffreys-Lindley paradox. Additionally, the BF lacks calibration, as its finite sampling distribution is unknown and may depend on nuisance parameters. To overcome these limitations, recent discussions have explored alternative Bayesian measures of evidence for testing precise null hypotheses, such as the $e-$value (see e.g. [3], and references therein) and the Bayesian Discrepancy Measure [1].

In particular, here we focus on the Bayesian Discrepancy Measure (BDM) of evidence proposed in [1]. The BDM gives an absolute evaluation of a hypothesis H_0 in light of prior knowledge about the parameter and observed data. In the absolutely continuous case, for testing (2) in the presence of nuisance parameters, the BDM is defined as

$$\delta_H = 1 - 2\min\left\{\int_{-\infty}^{\psi_0} \pi_m(\psi|y)\,d\psi, 1 - \int_{-\infty}^{\psi_0} \pi_m(\psi|y)\,d\psi\right\}. \tag{3}$$

The quantity $\min\{\int_{-\infty}^{\psi_0} \pi_m(\psi|y)\,d\psi, 1 - \int_{-\infty}^{\psi_0} \pi_m(\psi|y)\,d\psi\}$ can interpreted as the posterior probability of a "tail" event concerning only the precise hypothesis H_0. Doubling this "tail" probability, related to the precise hypothesis H_0, one gets a posterior probability assessment about how "central" the hypothesis H_0 is, and hence how it is supported by the prior and the data. This interpretation is related to an alternative definition for δ_H. Let ψ_m be the (marginal) posterior median and consider the interval defined as $I_E = (\psi_0, +\infty)$ if $\psi_m < \psi_0$ or as $I_E = (-\infty, \psi_0)$ if $\psi_0 < \psi_m$. Then, the BDM of the hypothesis H_0 can be computed as

$$\delta_H = 1 - 2\,P(\psi \in I_E|y) = 1 - 2\int_{I_E} \pi_m(\psi|y)\,d\psi. \tag{4}$$

Note that $2P(\psi \in I_E|y)$ gives the posterior probability of an equi-tailed credible interval for ψ.

The Bayesian Discrepancy Test (BDT) evaluates hypothesis H_0 using the Bayesian Discrepancy Measure. High values of δ_H indicate strong evidence against H_0, while small values suggest data consistency with H_0. When H_0 is true, for large sample sizes, δ_H is equally likely to fall between 0 and 1. Conversely, when H_0 is false, δ_H is more likely to approach 1 than 0. While thresholds can be set to interpret δ_H, in line with the ASA statement, we agree with Fisher that significance levels should be tailored to each case based on evidence and ideas.

The BDM remains invariant under invertible monotonic reparametrizations. Under general regularity conditions and assuming Cromwell's Rule for prior selection, δ_H exhibits specific properties: (1) if $\psi_0 = \psi_t$ (the true value), δ_H tends toward a uniform distribution as sample size increases; (2) if $\psi_0 \neq \psi_t$, δ_H converges to 1 in probability. In the case of a matching prior where $\psi_0 = \psi_t$, δ_H follows a uniform distribution between 0 and 1 for all sample sizes.

The practical computation of δ_H requires the evaluation of integrals of $\pi_m(\psi|y)$. The aim of this paper is to discuss higher-order expansions for δ_H

(Sect. 2). Moreover, connections with frequentist inference are highlighted when using objective priors (Sect. 3). Finally, an extension to the multivariate framework is provided (Sect. 4). Some final remarks are discussed in Sect. 5.

2 Asymptotic Approximations for the BDM

Let $\ell_p(\psi) = \log L(\psi, \hat{\lambda}_\psi)$ be the profile loglikelihood for ψ, with $\hat{\lambda}_\psi$ constrained maximum likelihood estimate (MLE) of λ given ψ. Moreover, let $(\hat{\psi}, \hat{\lambda})$ be the full MLE, and let $j_p(\psi) = -\partial^2 \ell_p(\psi)/\partial\psi^2$ be the profile observed information. A first-order approximation for δ_H when testing (2) is simply given by

$$\delta_H \doteq 2\,\Phi\left(\left|\frac{\psi_0 - \hat{\psi}}{\sqrt{j_p(\hat{\psi})^{-1}}}\right|\right) - 1, \tag{5}$$

where the symbol "$\doteq$" indicates that the approximation is accurate to $O(n^{-1/2})$ and $\Phi(\cdot)$ is the standard normal distribution function. Thus, to first-order, δ_H agrees numerically with $1-p$-value based on the profile Wald statistic $w_p(\psi) = (\hat{\psi} - \psi_0)/j_p(\hat{\psi})^{-1/2}$. In practice, the approximation (5) of δ_H is often inaccurate, in particular when the dimension of λ is large with respect to the sample size, because it forces the marginal posterior distribution to be symmetric.

In order to have more accurate evaluations of δ_H, it may be useful to resort to higher-order approximations based on tail area approximations (see, e.g., [5,9], and references therein). Using the tail area argument to the marginal posterior density, we can derive the $O(n^{-3/2})$ approximation

$$\int_{\psi_0}^{\infty} \pi_m(\psi|y)\, d\psi \doteqdot \Phi(r_B^*(\psi_0)), \tag{6}$$

where the symbol "$\doteqdot$" indicates that the approximation is accurate to $O(n^{-3/2})$ and

$$r_B^*(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)} \log \frac{q_B(\psi)}{r_p(\psi)},$$

with $r_p(\psi) = \text{sign}(\hat{\psi} - \psi)[2(\ell_p(\hat{\psi}) - \ell_p(\psi))]^{1/2}$ profile likelihood root and

$$q_B(\psi) = \ell_p'(\psi)|j_p(\hat{\psi})|^{-1/2} \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}} \frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi, \hat{\lambda}_\psi)}.$$

In the expression of $q_B(\psi)$, $\ell_p'(\psi)$ is the profile score function and $j_{\lambda\lambda}(\psi, \lambda)$ represents the (λ, λ)-block of the observed information $j(\psi, \lambda)$.

Using the tail area approximation (6), a third-order approximation of the BDM (4) can be computed as

$$\delta_H \doteqdot 1 - 2\min\{\Phi(r_B^*(\psi_0)), 1 - \Phi(r_B^*(\psi_0))\}. \tag{7}$$

Note that the higher-order approximation (7) does not call for any condition on the prior $\pi(\psi, \lambda)$, i.e. it can be also improper. Note also that using $r_B^*(\psi)$ an $(1-\alpha)$ equi-tailed credible interval for ψ can be computed as $CI = \{\psi : |r_B^*(\psi)| \leq z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution, and in practice it can reflect asymmetries of the posterior. Moreover, from (6), the posterior median of (1) can be computed as the solution in ψ of the estimating equation $r_B^*(\psi) = 0$.

3 Higher-Order Agreement Between Frequentist and Bayesian Evidence

Let us consider a so-called *strong* matching prior ([2,9,10]), i.e. a prior such that a frequentist p-value coincides with a Bayesian posterior survivor probability to a high degree of approximation, in the marginal posterior density (1). In this case, following [9], when using such matching prior, the marginal posterior density can be written as

$$\pi_m(\psi|y) \doteqdot \exp\left(-\frac{1}{2} r_p^*(\psi)^2\right) \left|\frac{s_p(\psi)}{r_p(\psi)}\right|, \tag{8}$$

where $s_p(\psi) = \ell_p'(\psi)/j_p(\hat{\psi})^{1/2}$ is the profile score statistic. Moreover, the tail area of the marginal posterior for ψ can be approximated to third-order as

$$\int_{\psi_0}^{\infty} \pi_m(\psi|y)\, d\psi \doteqdot \Phi(r_p^*(\psi_0)), \tag{9}$$

where $r_p^*(\psi)$ is the modified profile likelihood root

$$r_p^*(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)} \log \frac{q_p(\psi)}{r_p(\psi)}, \tag{10}$$

which has a third-order standard normal null distribution. In (10), the quantity $q_p(\psi)$ is a suitably defined correction term (see, e.g., [7], Chap. 9). Using (9), an asymptotic equi-tailed credible interval for ψ can be computed as $CI = \{\psi : |r_p^*(\psi)| \leq z_{1-\alpha/2}\}$, i.e., as a confidence interval for ψ based on (10) with approximate level $(1-\alpha)$. Note from (9) that the posterior median of $\pi_m(\psi|y)$ can be computed as the solution in ψ of the estimating equation $r_p^*(\psi) = 0$, and thus it coincides with the frequentist estimator defined as the zero-level confidence interval based on

Using the tail area approximation (9), a third-order approximation of the BDM of the hypothesis $H_0 : \psi = \psi_0$ is

$$\begin{aligned} \delta_H^* &= 2\min\{\Phi(r_p^*(\psi_0)), 1 - \Phi(r_p^*(\psi_0))\} - 1 \\ &= 2\,\Phi(|r_p^*(\psi_0)|) - 1. \end{aligned} \tag{11}$$

In this case (11) coincides with $1 - p_r^*$, where p_r^* is the p-value based on (10).

Moreover, note that using $r_p^*(\psi)$ an $(1-\alpha)$ the equi-tailed credible interval for ψ, given by $CI = \{\psi : |r_B^*(\psi)| \leq z_{1-\alpha/2}\}$, coincides with an accurate higher-order likelihood-based confidence interval for ψ. Finally, the posterior median coincides with the frequentist estimator defined as the zero-level confidence interval based on $r_p^*(\psi)$. Such an estimator has been shown to be a refinement of the MLE $\hat{\psi}$.

Thus, when using strong matching priors and higher-order asymptotics, there is an agreement between Bayesian and frequentist testing hypothesis, point and interval estimation.

From a practical point of view, the computation of (11) can be easily performed in practical problems using the `likelihoodAsy` package [4] of the statistical software `R`. In practice, the advantage of using this package is that it does not require the function $q_p(\psi)$ explicitly but it only requires the code for computing the log-likelihood function and for generating data from the assumed model. Some examples can be found in [6].

4 Approximations for Multidimensional Parameters

Approximations based on modifications of the log-likelihood ratios are available also for multidimensional parameters of interest, both with or without nuisance parameters (see [8–10], and references therein).

Paralleling the scalar parameter case, a credible region for a d-dimensional parameter of interest ψ with approximately $100(1-\alpha)\%$ coverage in repeated sampling, can be computed as

$$CR = \{\psi : w^\star(\psi) \leq \chi^2_{d;1-\alpha}\},$$

where $w^\star(\psi)$ is a suitable modification of the log-likelihood ratio or of the profile log-likelihood ratio (see [8,9]), and $\chi^2_{d;1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2_d distribution.

In practice, the region CR can be interpreted as the extension to the multidimensional case of the equi-tailed set CI, i.e. the region CR is computed as a multidimensional case of the set CI, based on the chi-squared approximation. As in the scalar case, the region CR can reflect departures from symmetry, with respect to the first order approximation based on the Wald statistic. Some examples can be found in [10].

Using $w^\star(\psi)$, a higher-order approximation of the BDM of the hypothesis $H_0 : \psi = \psi_0$ can be obtained as

$$\delta_H^* = 1 - P(\chi^2_d \geq w^\star(\psi_0)). \tag{12}$$

Examples and applications of (12) are under investigation.

5 Concluding Remarks

One appealing feature of approximations (6) and (9) is that they may routinely be applied in practical Bayesian inference, since they require little more than standard likelihood quantities for their implementation, and hence they may be available at little additional computational cost over simple first-order approximations. Although the approximations described in this paper are derived from asymptotic considerations, they perform extremely well in moderate or even small sample situations.

When using objective Bayesian procedures based on strong matching priors and higher-order asymptotics, there is an agreement between Bayesian and frequentist point and interval estimation, and also in significance measures. This is not true in general with the $e-$value as discussed in [6].

References

1. Bertolino, F., Manca, M., Musio, M., Racugno, W., Ventura, L.: A new Bayesian discrepancy measure. Stat. Meth. App. (2024). https://doi.org/10.1007/s10260-024-00745-1
2. Fraser, D.A.S., Reid, N.: Strong matching of frequentist and Bayesian parametric inference. J. Stat. Plan. Inference **103**, 263–285 (2002)
3. Pereira, C., Stern, J.M.: The $e-$value: a fully Bayesian significance measure for precise statistical hypotheses and its research program. Sao Paulo J. Math. Sci. **16**, 566–584 (2022)
4. Pierce, D.A., Bellio, R.: Modern likelihood-frequentist inference. Int. Stat. Rev. **85**, 519–541 (2017)
5. Reid, N.: The 2000 Wald memorial lectures: asymptotics and the theory of inference. Ann. Statist. **31**, 1695–1731 (2003)
6. Ruli, E., Ventura, L.: Can Bayesian, confidence distribution and frequentist inference agree? Stat. Meth. Ap. **30**, 359–373 (2021)
7. Severini, T.A.: Likelihood Methods in Statistics. Oxford University Press, Oxford (2000)
8. Skovgaard, I.M.: Likelihood asymptotics. Scand. J. Stat. **28**, 3–32 (2001)
9. Ventura, L., Reid, N.: Approximate Bayesian computation with modified loglikelihood ratios. Metron **7**, 231–245 (2014)
10. Ventura, L., Ruli, E., Racugno, W.: A note on approximate Bayesian credible sets based on modified log-likelihood ratios. Stat. Prob. Lett. **83**, 2467–2472 (2013)



A Latent Curve Model to Estimate the Evolution of Urban Air Pollution

Bottazzi Schenone Mariaelena[1(✉)] and Grimaccia Elena[2]

[1] Department of Statistical Sciences, Sapienza University, Rome, Italy
mariaelena.bottazzischenone@uniroma1.it
[2] ISTAT - Italian National Institute of Statistics, Rome, Italy
elgrimac@istat.it

Abstract. The goal of this study is to evaluate the evolution of the latent construct of air pollution, accounting for the interplay of time-varying covariates, specifically temperature variations, and geographical factors such as latitude and longitude. This is achieved employing a Latent Curve Model that quantifies and interprets the changes of the Air Pollution Index in Europe across the years 2019 to 2023. This analysis is crucial for discerning long-term trends in air pollution, providing valuable insights into potential environmental shifts that may impact public health and guide future policy considerations. Results show an increase of European air pollution level in the last five years. This trend is closely linked with increasing temperatures, underscoring the urgent need to address the intertwined challenges of air quality and climate change.

Keywords: Latent Curve Model · air pollution · climate change · European urban areas

1 Introduction

Monitoring air pollution is crucial as it plays an important role in safeguarding public health and the environment. In recent years, the impact of climate change has become evident, stressing even more the challenges associated with air quality [1]. Rising global temperatures, altered precipitation patterns, and changing weather conditions contribute to the intensification of air pollution. Using a statistical technique to model air pollution, we gain insights into the dynamic interplay between climate change and air quality, enabling the development of effective strategies to mitigate the adverse effects on the human health and the ecosystem. Indeed, comprehensive air quality studies not only enhance our understanding of environmental changes but also empower policy makers to implement targeted interventions to address the evolving challenges posed by climate-induced alterations in air composition. In this context, many different air pollution models have been proposed for short-term forecast, long-term projection, air-pollutant exposure and risk assessments, and most of them includes spatial and spatio-temporal effects. For instance, Taghavi et al. [2] use Functional Principal Component Analysis in the spatio-temporal land-use regression

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 266–272, 2025.
https://doi.org/10.1007/978-3-031-64431-3_45

modeling of Particulate Matter 2.5, while Ahmadi Basiri et al. [3] employ Functional Kriging for spatio-temporal modeling of Nitrogen Dioxide in Middle East. Multiscale geographical and temporal weighted Regression Methods are used by Yue et al. [4] and Li et al. [5] to model the determinants of Particulate Matter 2.5 in China, considering localized spatio-temporal effects. In this paper, Latent Curve Models (LCMs) are employed to analyse the dynamic trends in the multidimensional latent construct of air pollution. These flexible models can describe the evolution of air pollution over time, taking into account temperature changes as well as spatial effects.

This paper is structured as follows. Section 2 illustrates the data employed in the empirical study, while in Sect. 3 LCMs are theoretically presented. Section 4 is devoted to the air pollution analysis for 106 European urban areas between 2019 and 2023. Finally, in Sect. 5, concluding remarks are drawn and the key findings and insights derived from the empirical study are summarized.

2 Data

The data on which the study is conducted come from the Wordwide Air Quality database (https://aqicn.org/), and refer to the concentrations of the six main air pollutants, as identified by the Environmental Protection Agency: Ground-level ozone (O_3), Particle pollution (also known as Particulate Matter (PM), including $PM2.5$ and $PM10$), Carbon monoxide (CO), Sulfur dioxide (SO_2) and Nitrogen dioxide (NO_2) for 106 European urban areas for the years 2019 to 2023. For each urban area and year, the pollutants concentrations are computed as the average over the 365 d of the daily median values. These variables have been firstly employed to construct an Air Pollution Index (API) through Structural Equation Models (SEM), as in Bottazzi Schenone et al. [10]. This index ranges in $[0 - 1]$, where 1 corresponds to the highest pollution level, and 0 to the lowest, after a min-max normalization procedure. To quantify the impact of climatic changes on the Air Pollution Index, also temperature, humidity, wind gust, wind speed and pressure variables are considered for each urban area, from the same database. Latitude and longitude are also taken into account to control for spatial effects [6].

3 Methodology

Latent Curve Models are widely employed methodologies in longitudinal data analysis to examine the dynamic relationships among multiple variables over time. These models integrate latent variables, which represent unobservable constructs, with observed variables to describe underlying trends, patterns, and interdependencies that characterize the phenomenon under study. In our case, we will employ a latent curve to explore the evolution of the Air Pollution Index in Europe in the last five years.

Consider a population of N units observed at T times. A Latent Curve Model can be mathematically formalized as follows [7,8]. In a general notation, let y_{it} be

the observations of the response variable for unit i ($i = 1, \ldots, N$) at time t ($t = 1, \ldots, T$), α_i the intercept associated to the i-th unit ($\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_N$), β_i the slope associated to the i-th unit ($\boldsymbol{\beta} = \beta_1, \ldots, \beta_N$) and λ_t the time trend variable. Consider also J time-varying covariates, with w_{jit} denoting the j-th covariate for unit i at time t and γ_{jt} representing the corresponding coefficient. Let ϵ_{it} be the errors associated with the i-th unit at time t. The equation determining the trend of the response variable y_{it} is:

$$y_{it} = \alpha_i + \lambda_t \beta_i + \sum_{j=1}^{J} \gamma_{jt} w_{jit} + \epsilon_{it}. \tag{1}$$

A common convention for the temporal coefficients is to set, for each $t = 1, \ldots, T$, $\lambda_t = t - 1$ for linear trends. Other typical assumptions in regression analyses are that errors are centred ($\mathbb{E}(\epsilon_{it}) = 0 \ \forall i = 1, \ldots, N, \ t = 1, \ldots, T$) and normally distributed, and that α_i and β_i are uncorrelated with $\epsilon_{it} \ \forall i = 1, \ldots, N, \ t = 1, \ldots, T$. The equations describing the intercept and the slope are:

$$\begin{cases} \alpha_i & = \mu_{\boldsymbol{\alpha}} + \sum_{q=1}^{Q} \delta_{\alpha q} x_{qi} + \xi_{\alpha i} \\ [8pt]\beta_i & = \mu_{\boldsymbol{\beta}} + \sum_{q=1}^{Q} \delta_{\beta q} x_{qi} + \xi_{\beta i}, \end{cases} \tag{2}$$

where $\mu_{\boldsymbol{\alpha}}$ and $\mu_{\boldsymbol{\beta}}$ are the mean intercept and slope, respectively. A total of Q time-invariant covariates is considered in the model. The q-th time invariant covariate for the i-th unit is x_{qi}, with coefficients $\delta_{\boldsymbol{\alpha} q}$ and $\delta_{\boldsymbol{\beta} q}$. The residuals $\xi_{\alpha i}$ and $\xi_{\beta i}$ have zero mean are normally distributed and are uncorrelated with the ϵ_{it}. Estimates are obtained within the Maximum Likelihood framework [9].

4 Application

In this Section, a LCM is applied to air pollution, measured by the Air Pollution Index, considering 106 European urban areas from 2019 to 2023. The air pollution situation in Europe according to API is depicted in Fig. 1 for the year 2023.

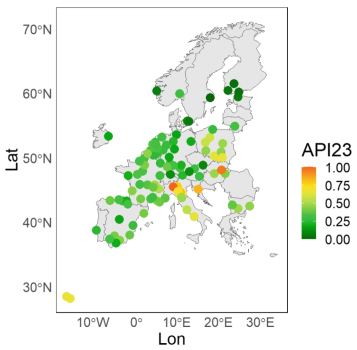


Fig. 1. Air Pollution Index values for European urban areas in 2023.

According to API, the most polluted areas, marked in orange and yellow, are located in Italy, Hungary, and Croatia. Notably, the two yellow points at the bottom-left of the figure are Tenerife and Las Palmas: these two urban areas were among the less polluted ones until 2021, while their air quality conditions have deeply worsened in the latest two years, due to the increasing population (see also Table 1).

Table 1 presents a ranking of European urban areas based on their API scores over the considered five-year period. Top ranks refer to the most polluted areas.

Table 1. Most (top ranks) and less (bottom ranks) polluted urban areas in Europe, according to the API ranking (2019–2023).

Rank	2019	2020	2021	2022	2023
1	Katowice	Milan	Modena	Milan	Milan
2	Milan	Parma	Katowice	Modena	Miskolc
3	Miskolc	Modena	Parma	Zagreb	Zagreb
4	Plovdiv	Miskolc	Miskolc	Brescia	Modena
5	Brescia	Plovdiv	Cracow	Parma	Brescia
...	...	...	...	...	...
102	Tenerife	Stockholm	Dublin	Stockholm	Malmo
103	Tallinn	Helsinki	Tenerife	Tampere	Turku
104	Tampere	Bergen	Santander	Turku	Helsinki
105	Las Palmas	Tallinn	Zurich	Helsinki	Tampere
106	Bergen	Dublin	Turku	Tallinn	Stockholm

Milan appears to have the highest rank for three years (2020,2022,2023), indicating a persistent air pollution issue. It is also possible to note a shift in the rankings of certain areas such as Tenerife, Las Palmas and Dublin, which were in the last positions until 2021, suggesting a recent change in their air quality conditions. Less polluted zones remain in Finland and Sweden across all the time period considered, while we can notice an increased number of Italian urban areas in the top ranking positions.

A LCM is applied considering the APIs scores as response variables ($N = 106$, $T = 5$), the meteorological variables as time-varying covariates ($J = 5$) and latitude and longitude as time-invariant covariates ($Q = 2$). From the complete model with all the covariates, we proceed with a backward selection, excluding at each step the non-statistically significant covariates. As a result, we selected the model shown in Fig. 2 ($J = 1, Q = 2$). The normality of model's residuals, computed as difference between the API score and the estimated one according to (1) for each year, has been assessed through the Shapiro-Wilks test.

The intercept coefficient $\mu_{\boldsymbol{\alpha}}$ corresponds to the average European API in 2019 and it is equal to 0.592, with a standard deviation of 0.015. The slope coefficient $\mu_{\boldsymbol{\beta}}$ captures the linear trend in the Air Pollution Index over the last five

years, and it is equal to 0.011, with a standard deviation of 0.003. This positive coefficient highlights a small, yet statistically significant increase in the European average API between 2019 and 2023. The first model's finding is, indeed, the significant growth of air pollution during the period in analysis. Furthermore, the positive relationship between $\boldsymbol{\beta}$ and temperature over time suggests a substantial influence of urban temperatures on escalating pollution levels. These findings not only confirm but also quantify what is known in literature [11] [12], offering valuable insights specific for the European context.

The model provides also the coefficients associated to temperature on API over time. The coefficients' estimates are positive, indicating that urban areas' temperature has a statistically significant effect on the increasing air pollution level. Moreover, the fact that they have strongly increased in the last two years warns about the impact that increasing temperatures due to climate changes may have on air pollution.

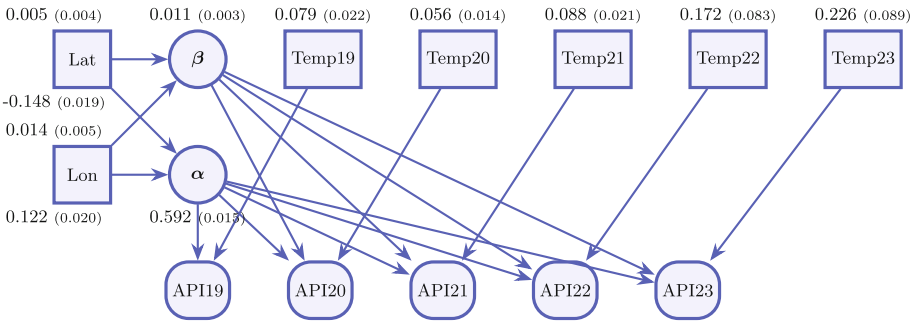


Fig. 2. Path diagram of the API LCM, estimates' standard errors are reported in parentheses.

Finally, the model reveals disparities within the European area. The positive longitude coefficient and negative latitude coefficient associated to the intercept show that Northern and Western Europe urban areas generally experience lower pollution levels. This observation is consistent with extant environmental research. Notably, the latitude's non statistically significant effect on the slope indicates that the pollution level divergence between Northern and Southern European urban areas has not widened over time: it may even suggest a slight convergence. Conversely, the significant positive coefficient for longitude points to a growing divergence, indicating that Eastern European urban areas are becoming increasingly more polluted than their Western counterparts.

5 Concluding Remarks

In this study, we present a Latent Curve Model tailored to explore the dynamics of multiple variables over time and space, focusing on the relationship between

air pollution and temperature. By adjusting for spatial covariates, our approach mitigates potential biases in the estimates that might arise if these factors were overlooked.

Through this work, we achieve several key findings. Primarily, the analysis reveals a concerning escalation in air pollution across Europe over the last five years, underscoring the urgency of addressing this environmental challenge. Additionally, the proposed model reveals a significant association between climatic changes and pollution levels, indicating a concurrent rise in temperature and contaminants' concentration. The model's capability to assess spatial covariates further enriches our understanding, showing an almost stable difference over time among Northern and Southern European areas and a slight increase of air pollution in Eastern Europe.

Future research could be devoted to identify and integrate additional covariates into the model, also considering non linear interactions among them. Such inclusions would not only enhance model's comprehensiveness and accuracy, but also improve its predictive capability. Furthermore, by adopting a more granular temporal resolution such as monthly intervals instead of yearly, it would be possible to gain more time-detailed insights into the phenomenon under study, addressing for instance seasonality effects.

References

1. Pinho Gomes, A.C., et al.: Air pollution and climate change. Lancet. Planetary Health **7**(9), e727–e728 (2023). https://doi.org/10.1016/S2542-5196(23)00189-4
2. Taghavi, M., et al.: Application of functional principal component analysis in the spatiotemporal land-use regression modeling of pm2.5. Atmosphere **14**(6), 926 (2023). https://doi.org/10.3390/atmos14060926
3. Ahmadi Basiri, E., Taghavi-Shahri, S.M., Mahaki, B., Amini, H.: Functional kriging for spatiotemporal modeling of nitrogen dioxide in a middle eastern megacity. Atmosphere **13**(7), 1095 (2022). https://doi.org/10.3390/atmos13071095
4. Yue, H., Duan, L., Lu, M., Huang, H., Zhang, X., Liu, H.: Modeling the determinants of pm2.5 in china considering the localized spatiotemporal effects: a multiscale geographically weighted regression method. Atmosphere **13**(4), 627 (2022). https://doi.org/10.3390/atmos13040627
5. Li, Q., Li, X., Li, H.: Factors influencing pm2.5 concentrations in the Beijing-Tianjin-Hebei Urban agglomeration using a geographical and temporal weighted regression model. Atmosphere **13**(3), 407 (2022). https://doi.org/10.3390/atmos13030407
6. Urdangarin, A., Goicoa, T., Ugarte, M. D.: Evaluating recent methods to overcome spatial confounding. Revista Matemática Complutense **36**, 333–360 (2022). https://doi.org/10.1007/s13163-022-00449-8
7. Bollen, K.A., Curran, P.J.: Latent Curve Models: A Structural Equation Perspective. Wiley Inc, New York (2006)
8. Allison, P.D.: Fixed Effects Regression Models. SAGE Publications Inc, New York (2009)
9. Meredith, W., Tisak, J.: Latent curve analysis. Psychometrika **55**(1), 107–122 (1990). https://doi.org/10.1007/BF02294746

10. Bottazzi Schenone, M., Grimaccia, E., Vichi, M.: Computational assessment of k-means clustering on a structural equation model based index. In: Book of Short Papers SIS 2023 - Scientific meeting of the Italian Statistical Society, pp. 978-983. Pearson, Milano (2023)
11. Jacob, D.J., Winner, D.A.: Effect of climate change on air quality. Atmos. Environ. **43**(1), 51–63 (2009). https://doi.org/10.1016/j.atmosenv.2008.09.051.
12. The Royal Society: Effects of net-zero policies and climate change on air quality. DES7687 (2021). ISBN: 978-1-78252-558-5

Measures of Interrater Agreement for Ratings on Nominal Scale

Giuseppe Bove[1(✉)] and Daniela Marella[2]

[1] Universitá Roma TRE, Dipartimento di Scienze della Formazione, Via del Castro Pretorio 20, 00185 Rome, Italy
giuseppe.bove@uniroma3.it

[2] Sapienza Universitá di Roma, Dipartimento di Scienze Sociali ed Economiche, Piazzale Aldo Moro 5,00185 Rome, Italy
daniela.marella@uniroma1.it

Abstract. The extent of agreement between raters on nominally scaled data regarding a group of targets is usually computed by kappa-type indices. Recently, the homogeneity index for a qualitative variable was proposed to evaluate agreement between raters for each target, and to obtain a global measure of interrater agreement for the whole group of targets. In this presentation, some sampling properties of the proposed indices are investigated and a comparative application regarding medical data is provided.

Keywords: nominal classification scales · interrater agreement · homogeneity index

1 Introduction

Applications of interrater agreement (IRA) indices for nominal scales are frequent in interventional and observational studies. For instance, information on disease status is often collected in a nominal way, like when clinical diagnoses provided by more physicians (raters) who recognize symptoms with respect to type of disease is considered for identifying the best treatment for patients (targets). High reliability of the rating procedure is crucial for providing meaningful and trustable treatments. For nominal data, a common approach for assessing IRA would be to compute the value of an index, and researchers are interested in assessing reliability at multiple levels ([7]). The most studied level is the overall agreement of the group of raters and the most common used measures are the kappa index or one of its generalizations for more than two raters (not necessarily the same raters for each target; for a review see [4]). A level considered much less than the overall agreement of the group of raters is the single target agreement (*e.g.*, [6]), in spite of the fact that it is particularly useful for improving the rating procedure (it is possible to request the raters a comparison on targets in which the agreement is poor and modify the procedure accordingly). In [1] a single target index and an overall index of IRA were proposed that present

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 273–278, 2025.
https://doi.org/10.1007/978-3-031-64431-3_46

advantages respect to the kappa-type indices and the O'Connell and Dobson single target index. The indices proposed were presented as measures of the size of the IRA and lack a statistical base for the implementation of confidence intervals or significance tests. In the next sections, after a brief presentation of the indices proposed in [1], some preliminary results regarding their sampling properties are presented. A final comparison with the most used indices of IRA for nominal scales is provided by a set of data regarding the cause of death of 35 hypertensive patients.

2 Measures of Interrater Agreement for Nominal Scales

In [6] a single target measure of interrater agreement is proposed. More specifically, for the ith target ($i = 1, \ldots, n_T$) the index is defined as follows,

$$S_i = 1 - \frac{D_i}{\Delta} \tag{1}$$

where D_i represents the overall disagreement and Δ the disagreement by chance, computed on the basis of the observed proportion of targets that each rater allocates to the category of the scales. A disagreement function for the target ith and a pair of raters is defined as 0 if raters agree and 1 if raters disagree, then the overall disagreement is obtained as the sum of disagreements on all $\binom{n_R}{2}$ possible pairs of the n_R raters.

An overall measure of agreement across targets can be obtained as the average of the S_i individual values $S_{av} = \sum_{i=1}^{n_T} S_i / n_T$.

An alternative approach is in [1] where a single target index of IRA is proposed. Such an index is based on the well-known homogeneity index O to measure the dispersion of a nominal variable having q categories (*e.g.*, [5]), given by

$$O = \sum_{k=1}^{q} \theta_k^2 \tag{2}$$

where θ_k is the proportion of ratings in category k, for $k = 1, \ldots, q$.

The index is equal to 1 in the case of maximum homogeneity (perfect agreement), and $1/q$ in the case of maximum heterogeneity (total disagreement, for each category k is $\theta_k = 1/q$). O depends on the number of categories, so the normalization in the interval $[0, 1]$ is given by

$$O_{rel} = \frac{qO - 1}{q - 1}. \tag{3}$$

It is not difficult to verify that O_{rel} is the normalized variance of the proportions θ_k ($k = 1, 2, \ldots, q$), and it is equal to 1 minus Gini's heterogeneity index (see also [3]).

3 Sampling Proprieties of O_{rel} Index

In our framework we assume that n_R raters rate n_T targets randomly drawn from the population of targets.

Let X be a nominal variable assuming q categories. Let us denote with X_{ij} the category that rater j assigns to target i, for $i = 1, \ldots, n_T$ and $j = 1, \ldots, n_R$. Formally, X_{ij} are independent categorical random variables having q categories with $\theta_k^{ij} = Prob(X_{ij} = k)$ for $i = 1, \ldots, n_T$, $j = 1, \ldots, n_R$ and $k = 1, \ldots, q$.

In the sequel we assume that both the targets and the raters are homogeneous (targets-raters homogeneity assumption), which implies that the probability $\theta_k^{ij} = \theta_k$ does not depend on rater j and on target i, for $i = 1, \ldots, n_T$, $j = 1, \ldots, n_R$ and $k = 1, \ldots, q$. As consequence of the homogeneity assumption, the variables X_{ij} are independent and identically distributed (*i.i.d.*, for short).

As previously stressed, the dispersion of a nominal variable can be measured by the index (3).

With regard to the ith target, let us denote with $A_{ik} = \{j : X_{ij} = k\}$ the set of raters who classified target i into category k and with r_{ik} its size, for $k = 1, \ldots, q$. Clearly, $\sum_{k=1}^{q} r_{ik} = n_R$ since all raters rate target i. The random variable $(r_{i1}, r_{i2}, \ldots, r_{iq})$ follows a multinomial distribution with parameters $(n_R, (\theta_1, \theta_2, \ldots, \theta_q))$.

In order to estimate (3), for each target i the following estimator can be defined

$$\widehat{O}_{rel,i} = \frac{q\widehat{O}_i - 1}{q - 1} \tag{4}$$

where $\widehat{O}_i = \sum_{k=1}^{q} \left(\frac{r_{ik}}{n_R}\right)^2$. Since $E(r_{ik}) = n_R\theta_k$, $V(r_{ik}) = n_R\theta_k(1 - \theta_k)$, for $k = 1, \ldots, q$, the expectation of the estimator (4) can be easily computed as shown in Proposition 1.

Proposition 1. *The estimator $\widehat{O}_{rel,i}$ has expectation given by*

$$E(\widehat{O}_{rel,i}) = O_{rel} + \frac{q}{q - 1} \sum_{k=1}^{q} \frac{\theta_k(1 - \theta_k)}{n_R}. \tag{5}$$

The proof of Proposition 1 is in [2]. From Proposition 1, an unbiased estimator of O_{rel} can be defined as follows

$$\widehat{O}^c_{rel,i} = \widehat{O}_{rel,i} - \frac{q}{q - 1} \frac{1}{n_R - 1} \sum_{k=1}^{q} \frac{r_{ik}}{n_R} \left(1 - \frac{r_{ik}}{n_R}\right). \tag{6}$$

As estimator of the homogeneity index O_{rel}, we consider

$$\widehat{O}^c_{rel} = \frac{1}{n_T} \sum_{i=1}^{n_T} \widehat{O}^c_{rel,i} \tag{7}$$

obtained averaging the n_T estimates $\widehat{O}^c_{rel,1}, \ldots, \widehat{O}^c_{rel,n_T}$. In the following proposition the sampling proprieties and the asymptotic distribution of $\widehat{O}^c_{rel}$ are analyzed for large n_T.

Proposition 2. *The estimator $\widehat{O}^c_{rel}$ has expectation*

$$E(\widehat{O}^c_{rel}) = O_{rel}, \tag{8}$$

and variance

$$Var(\widehat{O}^c_{rel}) = \sum_{i=1}^{n_T} \frac{Var(\widehat{O}^c_{rel,i})}{n_T} = \frac{V}{n_T}. \tag{9}$$

where $V = Var(\widehat{O}^c_{rel,i})$. Furthermore, since $\widehat{O}^c_{rel,1}, \ldots, \widehat{O}^c_{rel,n_T}$ are i.i.d., for the central limit theorem, as n_T goes to infinity the random variable $\widehat{O}^c_{rel}$ tends to a standard normal distribution with mean and variance given by (8) and (9), respectively.

In order to estimate the variance of the estimator $\widehat{O}^c_{rel}$, a nonparametric as well as a parametric bootstrap procedure can be used. In the first case, each bootstrap sample is generated by selecting from the original sample a simple random sampling with replacement of n_T targets.

In the parametric bootstrap each bootstrap sample is generated according to a multinomial model with parameters $(n_R, (\bar{\widehat{\theta}}_1, \ldots, \bar{\widehat{\theta}}_q))$ where $\bar{\widehat{\theta}}_k = \frac{1}{n_T} \sum_{i=1}^{n_T} \widehat{\theta}_{k,i}$, and $\widehat{\theta}_{k,i} = r_{ik}/n_R$ is the estimate of the parameter θ_k obtained from the ith row of the initial sample, for $i = 1, \ldots, n_T$ and $k = 1, \ldots, q$.

The results in Proposition 2 are useful to construct interval estimates and to test hypothesis regarding O_{rel}.

4 Application

In [8] a study with seven nosologists ($n_R = 7$) assessing the cause of death of 35 hypertensive patients by using their death certificates is considered. The scores were assigned according to the following categories ($q = 5$): 1 = *Arteriosclerotic disease*, 2 = *cerebrovascular disease*, 3 = *other heart disease*, 4 = *renal disease*, and 5 = *other disease*. The marginal proportions of ratings for the five categories were $0.21, 0.17, 0.19, 0.27, 0.16$, respectively.

The biased estimator $\widehat{O}_{rel} = \sum_{i=1}^{n_T} \widehat{O}_{rel,i}/n_T = 0.56$, the unbiased estimator (7) is $\widehat{O}^c_{rel} = 0.49$. The first, the second and the third quartiles of the subjective values of $\widehat{O}^c_{rel,i}$ are $0.31, 0.40, 0.64$, which calls for a possible revision of the assessment procedure since 50% of the $\widehat{O}^c_{rel,i}$ values are smaller than or equal to 0.40.

In Table 1 some descriptive statistics are reported for the biased estimator $\widehat{O}_{rel,i}$, the unbiased estimator $\widehat{O}^c_{rel,i}$ and the index S_i proposed in [6].

Table 1. Some descriptive statistics for $\widehat{O}_{rel,i}$, $\widehat{O}^c_{rel,i}$ and S_i

Index	Mean	St.Dev.
S_i	0.48	0.27
$\widehat{O}_{rel\,i}$	0.56	0.24
$\widehat{O}^c_{rel,i}$	0.49	0.27

As shown in Table 1 the mean and the standard deviation of the unbiased estimator $\widehat{O}^c_{rel,i}$ and of the index S_i are approximately the same. The value of S_{av} is 0.48. So, both the indices S_{av} and $\widehat{O}^c_{rel}$ show a moderate level of agreement between the seven nosologists.

Finally, in order to construct confidence interval for O_{rel} the nonparametric bootstrap is used. Formally, B = 1000 bootstrap samples are generated by selecting a simple random sampling with replacement of $n_T = 35$ targets from the original sample of targets. More specifically, the following procedure is applied:

Step 1 For each bootstrap sample b (for $b = 1, \ldots, B$) compute the homogeneity index $\widehat{O}^c_{rel,b}$ and compute the mean of $\overline{\widehat{O}}^c_{rel} = (1/B)\sum_{b=1}^{B} \widehat{O}^c_{rel,b}$ over the bootstrap samples.

Step 2 Compute the variance of $\widehat{O}^c_{rel}$ as follows:

$$Var(\widehat{O}^c_{rel}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{O}^c_{rel,b} - \overline{\widehat{O}}^c_{rel})^2 \tag{10}$$

In Table 2, it is possible to compare the $(1 - \alpha = 0.95)$ confidence interval obtained for O_{rel} by using the normal approximation with the confidence intervals for S_{av} and some of the most used indices of IRA (see [4] for their definitions).

Table 2. Confidence intervals

Index	Value	St.Error	Confidence Interval
O^c_{rel}	0.488	0.046	(0.397, 0.580)
$S_{av}(hetero)$	0.483	0.047	(0.394, 0.575)
$S_{av}(homoge)$	0.477	0.048	(0.385, 0.572)
$Gwet'sAC1$	0.484	0.047	(0.389, 0.579)
$FleissKappa$	0.477	0.049	(0.378, 0.577)
$Krippendorf\alpha$	0.480	0.048	(0.383, 0.577)
$Brennan - Prediger$	0.402	0.045	(0.310, 0.495)

The coefficient values and their confidence intervals are almost coincident with the only exception of the Brennan-Prediger coefficient that presents an

higher disagreement respect to the other coefficients. $\widehat{O}^c_{rel}$ and S_{av} have the advantage to have the corresponding single target versions ($\widehat{O}^c_{rel,i}$ and S_i) useful for improving the rating procedure. Besides, $\widehat{O}^c_{rel,i}$ can be computed when we have observed only one target or when targets are rated by disjoint set of raters (*e.g.*, teachers rated by their classes), and it does not depend by the marginal distributions of the categories observed for each rater like most of the chance-corrected measure of IRA.

References

1. Bove, G.: The homogeneity index as a measure of interrater agreement for ratings on a nominale scale. In L.M. Grilli L. et al. (eds.), Statistical Models and Methods for Data Science, pp. 15–26. Cham : Springer Nature (2023)
2. Marella, D., Bove, G.: Statistical inference for measures of interrater agreement of ratings on a nominal scale. Submitted (2024)
3. Capecchi, S., Iannario, M.: Gini heterogeneity index for detecting uncertainty in ordinal data surveys. METRON **74**(2), 223–232 (2016). https://doi.org/10.1007/s40300-016-0088-5
4. Gwet, K.L.: Handbook of inter-rater reliability (4-th ed.), Gaithersburg MD, USA, Advanced Analytics, LLC (2014)
5. Leti, G.: Statistica Descrittiva. Il Mulino, Bologna (1983)
6. O'Connell, D.L., Dobson, A.J.: General observer-agreement measures on individual subjects and groups of subjects. Biometrics **40**, 973–983 (1984)
7. Posner, K.L., Sampson, P.D., Caplan, R.A., Ward, R.J., Cheney, F.W.: Measuring inter-rater reliability among multiple raters: an example of methods for nominal data. Stat. Med. **9**, 1103–1115 (1990)
8. Woolson, R.F.: Statistical Methods for the Analysis of Biomedical Data. Wiley, New York (1987)



Median-Based Inequality Measures

Vytaras Brazauskas[1], Francesca Greselin[2(✉)], and Ričardas Zitikis[3]

[1] University of Wisconsin-Milwaukee, Milwaukee, WI, USA
vytaras@uwm.edu
[2] Università degli Studi di Milano-Bicocca, Milano, Italy
francesca.greselin@unimib.it
[3] Western University, London, ON, Canada
rzitikis@uwo.ca

Abstract. We address income inequality, particularly in heavily skewed distributions where traditional measures like the Gini index - based on the population mean - may be inadequate. We explore three median-based inequality indices that reflect three views of measuring income inequality: in the first, the median income of the poor is compared with the median income of the entire population, in the second the comparison is made with the median income of those who are not poor, in the third the comparison is with the median of the same proportion of the richest. Several properties of the indices have been derived and discussed. The indices offer many advantages, the main one is that they are applicable to population distributions with any degree of tail heaviness, even not having any finite moment. An application to capital income data from European countries demonstrate the interpretation and the efficacy of the indices.

Keywords: measures of inequality · heavy-tailed distributions · income transfers

1 Introduction

Assessing income inequality has proven challenging because each index employed for this purpose seeks to distill into a single value the intricacies of populations. A wide literature exists on the subject, including treatises as Atkinson and Bourguignon's handbook [1], which extensively cites earlier studies. Additionally, insightful monographs have been authored along 100 years of research, from Gini [7] to Piketty [12], to cite the oldest and perhaps the more recent, among the many relevant ones.

In general, many classical inequality indices are the areas below the corresponding inequality curves. The curves can be obtained from comparing the mean income of the poor with various other means, including those of the entire population, the non-poor, and the rich. Observing the current socio-economic situation, the mean is not any more the best point of reference, because income

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 279–284, 2025.
https://doi.org/10.1007/978-3-031-64431-3_47

distributions evolve toward an increase in skewness and heavy-tailedness. Further, when dealing with capital income distributions, for some samples the Hill estimator of the tail index may take values into the interval (0.5,1), therefore the mean does not exist [9]. This discussion motivated us to shift the focus on the median, along the lines of Gastwirth [6], where the author proposed to normalize the Gini mean difference by the median to capture the recent evolutions occurring in income distributions. Here, we move a step forward by introducing three percentile-based indices of inequality, and their inequality curves.

1.1 Quantile-Based Indices for Current Income Distributions

We begin establishing briefly some notations. Let us denote by X a random variable describing the population incomes, and let F be its cumulative distribution function. We assume that F is non-negatively supported, that is, $F(x) = 0$ for all real $x < 0$. Furthermore, let Q denote the (generalized) inverse of F, called the quantile function. That is, for each $p \in (0, 1)$, $Q(p)$ is the smallest number x such that $F(x) \geq p$. Hence, the population median income is $m = Q(1/2)$, assumed throughout this paper to be strictly greater than 0. Generally, $Q(p)$ is the $p \times 100^{\text{th}}$ percentile. Furthermore, the median income of the poorest $p \times 100\%$ persons is $Q(p/2)$.

Now, coming back to the classical inequality indices, we start from the most widely employed one: the Gini index. It can be expressed through its inequality curve, which compares the mean income of the poor - those below the percentile $Q(p)$ - to the overall mean income [8]. A second popular index is the Zenga index [15]. It is based on contrasting the mean income of the poor, to the mean income of the remaining part of the population, being $Q(p)$ the threshold for such comparisons. Recently, Davydov and Greselin [4] looking at the extreme $p\%$ individuals of the population, suggested to compare the mean income of those below the percentile $Q(p)$ with the mean income of those above the percentile $Q(1-p)$.

In the following, we will traspose the same philosophy into three purely median-based indices of income inequality, and define - at the same time - the corresponding curves of equality. They reflect three different points of view, or perspectives. They are all based on percentiles and thus well defined, irrespective of whether the income variable X has a finite first or any other moment.

Perspective 1 Compare the median income of the poorest $p \times 100\%$ persons with the median of the entire population. This leads to the equality curve

$$\psi_1(p) = \frac{Q(p/2)}{Q(1/2)}, \quad 0 < p < 1, \tag{1}$$

also independently introduced by [10, equation (9)]. Averaging this curve over all p's gives rise to the inequality index

$$\Psi_1 = 1 - \int_0^1 \frac{Q(p/2)}{Q(1/2)} \mathrm{d}p. \tag{2}$$

Compare it with the left-most integral of [13, equation (3)]. Note the mathematical similarity between the Bonferroni curve b and the curve ψ_1:

$$b(p) = \frac{\frac{1}{p}\int_0^p Q(s)\mathrm{d}s}{\int_0^1 Q(s)\mathrm{d}s}, \quad \psi_1(p) = \frac{\frac{1}{p}\int_0^p Q(p/2)\mathrm{d}s}{\int_0^1 Q(1/2)\mathrm{d}s}.$$

Perspective 2 Compare the median income of the poorest $p \times 100\%$ persons with the median of the nonpoor. This leads to the equality curve

$$\psi_2(p) = \frac{Q(p/2)}{Q(1/2+p/2)}, \quad 0 < p < 1, \tag{3}$$

also independently introduced by [10, equation (10)]. Averaging this curve over all p's gives rise to the inequality index

$$\Psi_2 = 1 - \int_0^1 \frac{Q(p/2)}{Q(1/2+p/2)}\mathrm{d}p, \tag{4}$$

which is also considered by [11]. Note the mathematical similarity between the Zenga curve z [15] and the curve ψ_2:

$$z(p) = \frac{\frac{1}{p}\int_0^p Q(s)\mathrm{d}s}{\frac{1}{1-p}\int_p^1 Q(s)\mathrm{d}s}, \quad \psi_2(p) = \frac{\frac{1}{p}\int_0^p Q(p/2)\mathrm{d}s}{\frac{1}{1-p}\int_p^1 Q(p+(1-p)/2)\mathrm{d}s}.$$

Perspective 3 Compare the median income of the poorest $p \times 100\%$ persons with the median of the richest $p \times 100\%$ persons. This leads to the equality curve [14]

$$\psi_3(p) = \frac{Q(p/2)}{Q(1-p/2)}, \quad 0 < p < 1, \tag{5}$$

also considered by [10, equation (11)]. Compare this curve also with $L_2(F;p)$ of [13, Definition 1]. Averaging this curve over all p's gives rise to the inequality index [14]

$$\Psi_3 = 1 - \int_0^1 \frac{Q(p/2)}{Q(1-p/2)}\mathrm{d}p, \tag{6}$$

which is also considered by [11]. Note the mathematical similarity between the Davydov-Greselin curve d [5] and the curve ψ_3:

$$d(p) = \frac{\frac{1}{p}\int_0^p Q(s)\mathrm{d}s}{\frac{1}{p}\int_{1-p}^1 Q(s)\mathrm{d}s}, \quad \psi_3(p) = \frac{\frac{1}{p}\int_0^p Q(p/2)\mathrm{d}s}{\frac{1}{p}\int_{1-p}^1 Q(1-p+p/2)\mathrm{d}s}.$$

We developed a parametric viewpoint on the three curves and indices, using nine parametric families of distributions, all right skewed and with different degrees of tail heaviness. The analytical study cannot be incorporated here, for sake of space, but it can be found in our technical report [2].

1.2 Empirical Estimators for Quantile-Based Indices

Let $X_1, \ldots, X_n$ denote incomes of randomly selected persons, with $X_{1:n} \leq \cdots \leq X_{n:n}$ denoting the ordered incomes. We assume that the empirical median $Q_n(1/2) = X_{\lceil n/2 \rceil:n}$ is strictly positive, where, for every real $x \geq 0$, $\lceil x \rceil$ is the smallest integer that is not below x. The empirical counterparts of the three indices Ψ_k are (for their justifications, see the Appendix in [2]):

$$\Psi_{1,n} = 1 - \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{X_{k:n}}{X_{\lceil n/2 \rceil:n}}, \tag{7}$$

$$\Psi_{2,n} = 1 - \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{X_{k:n}}{X_{\lceil n/2 \rceil+k:n}}, \tag{8}$$

$$\Psi_{3,n} = 1 - \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{X_{k:n}}{X_{n-k+1:n}}, \tag{9}$$

where, for every real $x \geq 0$, $\lfloor x \rfloor$ is the largest integer that does not exceed x. Note that all the three indices are well defined because we assume that the median $X_{\lceil n/2 \rceil:n}$ is strictly positive.

It can be shown that the three indices enjoy the properties canonically required for inequality indices (with a wider interpretation for the Pigou-Dalton principle).

They are scale invariant, i.e. for every real $c \geq 0$, it holds $\Psi_{k,n}(c\mathbf{X}) = \Psi_{k,n}(\mathbf{X})$.

Further, except the case of all X's being equal, we have that $\Psi_{k,n}(\mathbf{X}) > \Psi_{k,n}(\mathbf{X} + c)$ for every real $c > 0$.

1.3 An Application to Capital Incomes from the EU-SILC2018 Survey

To get an insight into a recent European situation, we analyse data coming from the EU Statistics on Income and Living Conditions (EU-SILC2018) survey [3]. We focus on capital incomes, summing up the following variables across each household: HY040G: income from rental of a property or land; HY090G: interests, dividends, profit from capital investments in unincorporated business; and PY080: pensions received from individual private plans, so yielding the total of capital incomes $Z_1, \ldots, Z_n$ for the 340 540 surveyed households, from the fifteen European countries specified in Table 1. An equivalence scale must be utilized to facilitate meaningful comparisons of monetary incomes and incorporate economies of scale among social units with varying numbers of residents. We opt for the "square root" equivalence scale, adopted by the Organization for Economic Cooperation and Development (OECD) in their recent publications. Namely, each household income Z_i is divided by the square root of the household size w_i, obtaining X_i, $i = 1, \ldots, n$, that we use in our analysis. By looking at the means and medians in Table 1, we see how skewed to the right the distributions of the countries are. Based on empirical evidence drawn from the samples, [9]

show that some underlying income distributions lack finite first moment, so that classical indices could not be employed.

Portugal has the highest inequality in capital incomes, as measured by $\Psi_{1,n}$, with the median income of the poorest $p \times 100\%$ persons equal, after averaging over all $p \in (0,1)$, to 53% of the median income of the entire population. When we are interested in comparing the median income of the poorest $p \times 100\%$ persons with the median income of the remaining $(1-p) \times 100\%$ part of the population, the index $\Psi_{2,n}$ tells us that Finland is the country in which such a contrast, after averaging over all $p \in (0,1)$, is close to unity. Finland confirm to be the country with lowest inequality also when adopting Perspective 3, that is when we compare the median income of the poorest $p \times 100\%$ persons with the median income of the richest $p \times 100\%$ persons: after averaging such ratios over all $p \in (0,1)$, we obtain 90%.

Table 1. The Gini, Zenga and Davydov-Greselin income-inequality indices G_n, Z_n, D_n, the Gastwirth index $G_{2,n}$ [6], and the indices $\Psi_{1,n}$, $\Psi_{2,n}$, $\Psi_{3,n}$ for the fifteen European countries with $n = n_P$, where n_P is the number of people in the sample who possess capital incomes, and n_T is the total sample size of the given country based on EU-SILC2018.

Countries	Means	Medians	Sample sizes		Inequality indices							Ranks based on		
	$(in1000$	$Euro)$	n_T	n_P	G_n	Z_n	D_n	$G_{2,n}$	$\Psi_{1,n}$	$\Psi_{2,n}$	$\Psi_{3,n}$	$\Psi_{1,n}$	$\Psi_{2,n}$	$\Psi_{3,n}$
DE	1,516	147	25,784	20,332	0.85	0.92	0.98	8.71	0.55	0.94	0.76	3	4	3
DK	692	84	16,812	5,118	0.85	0.92	0.98	7.00	0.75	0.98	0.86	12	10	9
NL	1,542	85	24,986	19,192	0.91	0.96	0.99	16.52	0.64	0.96	0.81	6	5	5
BE	1,833	54	11,892	6,568	0.87	0.94	0.99	29.86	0.70	0.96	0.88	9	6	13
LU	3,193	125	7,666	4,400	0.85	0.93	0.99	21.88	0.63	0.97	0.82	5	9	7
FR	4,301	453	21,752	17,828	0.85	0.92	0.98	8.05	0.75	0.98	0.86	11	11	10
GB	2,811	442	34,226	15,090	0.79	0.89	0.97	5.01	0.70	0.96	0.84	8	8	8
IE	4,653	1,080	8,764	1,678	0.75	0.88	0.97	3.25	0.82	0.98	0.89	15	13	14
IT	2,004	267	42,346	22,188	0.81	0.90	0.98	6.08	0.66	0.96	0.82	7	7	6
GR	3,217	1,967	48,610	7,512	0.58	0.78	0.89	0.95	0.60	0.87	0.71	4	1	2
ES	2,132	264	26,736	13,246	0.81	0.90	0.98	6.50	0.74	0.98	0.87	10	12	11
PT	2,266	694	27,434	5,516	0.70	0.85	0.94	2.29	0.53	0.91	0.71	1	2	1
AT	1,699	104	12,206	8,598	0.88	0.94	0.99	14.36	0.54	0.93	0.77	2	3	4
FI	3,525	203	19,664	16,008	0.85	0.93	0.99	14.831	0.81	0.99	0.90	14	15	15
SE	313	33	11,662	9,138	0.84	0.92	0.98	7.88	0.76	0.99	0.87	13	14	12

1.4 Conclusions

The paper explores three inequality indices reflecting different approaches to measuring income inequality. Index Ψ_1 compares the median income of the poor to that of the entire population. Index Ψ_2 compares the median income of the poor to that of non-poor individuals. Index Ψ_3 compares the median

income of the poor to the median income of the same proportion of the richest. The indices under consideration can analyze populations without necessitating finite moments, making them applicable even to ultra-heavily tailed distributions. Unlike traditional indices like the Gini, which require a finite first moment, these indices offer flexibility. In such scenarios, statistical inference often leans on Extreme Value Theory. The application to capital income data from EU-SILC2018, where some underlying distributions do not possess first finite moment, show the value of the present proposal. Moreover, the indices enjoy the properties required for inequality measurements, and with respect to the classical Pigou-Dalton principle they tell us a richer story (for more details, we refer to [2]).

References

1. Atkinson, A.B. and Bourguignon, B.: Handbook of Income Distribution. (Volume 2.) Elsevier, Amsterdam (2015)
2. Brazauskas, V., Greselin, F. and Zitikis, R.: Measuring income inequality via percentile relativities. Technical report available at. (2023) https://doi.org/10.48550/arXiv.2308.03708
3. EU-SILC. EU Statistics on Income and Living Conditions. Eurostat, European Union. (2018). https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions
4. Davydov, Y., Greselin, F.: Inferential results for a new measure of inequality. Econometrics J. **22**, 153–172 (2019)
5. Davydov, Y., Greselin, F.: Comparisons between poorest and richest to measure inequality. Soc. Methods Res. **49**, 526–561 (2020)
6. Gastwirth, J.L.: Median-based measures of inequality: reassessing the increase in income inequality in the U.S. and Sweden. Stat. J. IAOS **30**, 311–320 (2014)
7. Gini, C.: Variabilità e Mutabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. Tipografia di Paolo Cuppini, Bologna (1912)
8. Greselin, F.: More equal and poorer, or richer but more unequal? Econ. Qual. Control **29**(2), 99–117 (2014)
9. Greselin, F., Pasquazzi, L., Zitikis, R.: Heavy tailed capital incomes: Zenga index, statistical inference, and ECHP data analysis. extremes: statistical theory and applications in science. Eng. Econ. **17**, 127–155 (2014)
10. Jokiel-Rokita, A. and Piątek, S.: Nonparametric estimators of inequality curves and inequality measures. Technical report available at (2023). https://doi.org/10.48550/arXiv.2310.16200
11. Piątek, S.: Parametric estimation of quantile versions of Zenga and D inequality curves: methodology and application to Weibull distribution. Technical report available at (2023). https://doi.org/10.48550/arXiv.2312.07350
12. Piketty, T.: Capital in the Twenty-First Century. English edition) Harvard University Press, Cambridge, Massachusetts (2014)
13. Prendergast, L.A., Staudte, R.G.: Quantile versions of the Lorenz curve. Electron. J. Stat. **10**, 1896–1926 (2016)
14. Prendergast, L.A., Staudte, R.G.: A simple and effective inequality measure. Am. Stat. **72**, 328–343 (2018)
15. Zenga, M.: Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. Stat. Appl. **5**, 3–27 (2007)

Improving Psychological Digital Wellbeing in Young Adult Users: A Bayesian Networks Approach

Chiara Brombin[1,2(✉)], Clelia Di Serio[1,2], and Paola Vicard[3]

[1] University Center for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, Via Olgettina 58, 20132 Milano, Italy
[2] Faculty of Psychology, Vita-Salute San Raffaele University, Via Olgettina 58, 20132 Milano, Italy
{brombin.chiara,diserio.clelia}@unisr.it
[3] Department of Economics,Università Roma Tre, Via Silvio D'Amico 77, 00145 Roma, Italy

Abstract. The digital revolution has not only transformed how people interact but also altered their relationship with self-image. The widespread availability of editing tools that enhance and refine self-portraits, allowing for an overall reconstruction of the face, has sparked concerns among experts as they promote unrealistic beauty standards. Different types and patterns of use on social networking sites (SNSs) may have varying impacts on users' psychological well-being, particularly when there is limited awareness of risks and digital skills. A Bayesian networks (BNs) framework is implemented to examine the intricate connections among SNSs usage, selfie-taking/sharing/editing behaviors, the digital self-image, internalizing symptoms, and other psychological aspects. In this setting, data dependence structure is learnt from data allowing for studying how changes in one or more variables propagates to all other variables so that alternative hypothetical scenarios may be explored. This presents an attractive statistical tool that supports the decision-making process regarding regulations on SNSs.

Keywords: Bayesian network · data-driven approach · digital wellbeing · propagation analysis · Social Networking Sites usage

1 Introduction

The diffusion of social networking sites (SNSs) and their intrinsic capability to rapidly adapt in response to both the technological advances and changing users needs and behaviors has significantly reshaped our daily life, opening from one side to new unexplored forms of communications and social relationships but also transforming our self-image representation in relation to our digital representation. Progresses achieved in the past years in face detection and recognition have been exploited also to develop photo editing tools and beauty filters for

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 285–290, 2025.
https://doi.org/10.1007/978-3-031-64431-3_48

self-portraits enhancement. Recently developed filters, capitalizing on artificial intelligence and augmented reality techniques, enable a real-time reconstruction of our face, globally altering facial appearance to the point that digital technology experts and health professionals are worried for users psychological wellbeing, especially relatively to younger users, in their developmental phase, and those more vulnerable. The misalignment between the high availability of these editing/digital tools and the difficulty of keeping up with digital technologies advances, coupled with difficulties in implementing effective media-education strategies for a responsible use of digital technologies, represents a challenge also for policymakers. In this setting, where also users behaviours are highly heterogeneous, thoroughly banning cannot be a solution. Through an online survey, the present study aims at exploring the use of SNSs, also in its different types, the selfie-taking/sharing/editing behavior, the relationship with digital self-image, the satisfaction with facial appearance, the perceived social and psychological function of the face, the role of internalizing symptoms (anxiety and depression) and appearance anxiety, focusing on young adults as target group. Differently from previous works emphasising body image [1], the relationship with the self-image is interpreted focusing mainly on the face. To examine and disentangling the complex interrelationships among collected variables and psychological constructs, gaining a better insight into mechanisms underlying self-(re)presentation in real and digital context, a Bayesian networks (BNs; [2]) framework is adopted. In this context, dependence structure is learnt directly from data allowing for "blacklists", hence forbidding some unfounded links among variables. Enjoying appealing statistical properties and high flexibility, BNs are widely and successfully used in various research areas, including socio-economic (e.g., see [3,4]), psychological (e.g., see [5]) and clinical (e.g., see [6]) fields. We exploit BNs computationally efficient inference engine to examine alternative hypothetical scenarios performing a "what-if" analysis, by analyzing how changes in one or more variables propagates to all other variables throughout the estimated network. BNs are appealing and effective statistical tools effectively used to provide evidences and support socialization agents (parents, teachers, health professionals) and policymakers when taking decisions about SNSs usage regulations.

2 SatisFace Project

Promoted by researchers from CUSSB (University Centre for Statistics in the Biomedical Sciences) and the Faculty of Psychology of Vita-Salute San Raffaele University (UniSR), the research project encompasses topics related to the field of social and developmental psychology, education, health and communication, by capitalizing on rigorous quali/quantitative statistical methodology. The project aims at investigating whether and how the SNSs (dys)functional use, socialization agents and psychological symptoms shape the digital-self construction/redefinition in a virtual space. Three research lines are currently active: one focusing on (pre)adolescents as target group, one focusing on healthy participants from the general population and chemotherapy-treated patients with

cancer diagnosis, and one focusing on University students jointly with Roma Tre University. Data on attitudes and behaviours are collected through online surveys consisting of both ad-hoc developed questions and validated psychometric tools. The study is exploratory in nature and relationships are mainly evaluated within a data-driven perspective. The project aims at providing a comprehensive socio-psychological characterization of SNSs users, identifying natural clusters of participants in the data with peculiar characteristics, while simultaneously targeting pivotal constructs on which leverage to enhance online behaviours. The applied methodological approach should contribute to enhance psychological knowledge on the impact of SNSs use generating new hypotheses to test in successive studies.

3 Sample Description

184 young adults (age ranging from 18–30yr, 68.5% females) took part to the online survey. Participants were mainly students (44%) or employed (32.6%), hailing from the northern and central regions of Italy (57.1% and 27.7% respectively). With respect to their education, 31.5% completed their secondary education, 39.7% obtained a bachelor degree, while 21.2% holds a master degree. 32.1% of participants reported to be highly skilled in the use of digital technologies, while 49.5% self-describes as moderately skilled (no difference between males and females). The most popular SNSs is Instagram (94.4% females, 94.8% males), followed by YouTube (31.7% females, 70.7% males) and TikTok (46.8% females, 32.8% males). 20.1% reported an early use of SNSs (i.e., they downloaded the first social media before 13yr). With reference to daily SNSs usage, 31.5% reported to spend from 2 to 4 h daily on Instagram (35.7% females, 22.4% males). 10.3% males (vs. 4.0% females) spend from 2 to 4 h daily on YouTube and 19% females (vs. 8.6% males) on TikTok. When investigating social media usage patterns, only females, but in a small percentage (1.6% of them), report to mainly create and actively share contents. The majority, 64.7%, report to passively use them, primarily consuming content shared by others (61.9% females, 70.7% males). Finally, 34.2% of participants report to be equally engaged in both active and passive interactions on social media platforms (36.5% females, 29.3% males). While reflecting from one side heterogeneity in online behaviors, combining the information on the time spent online with the reported ways of interacting on social media, it emerges a preference for passive interactions. Only 14.1% of participants report to mainly share permanent post or contents. About 47.3% (40.5% females, 62.1% males) reports to have 300–600 followers on average, with 43.7% females (vs. 20.7%) reporting to have more than 900 followers. 34.2% report to have a negative or very negative relationship with their face image reproduced in photos or videos (38.1% females, 25.9% males) When taking selfies, 32.6% reported to edit their selfies, adjusting image brightness or using filters (41.3% females, 13.8% males) mainly to enhance aesthetic qualities, feel more confident, appear as they would like hiding imperfection and to a lesser extent for obtaining more likes or because it is out of habit. Among those

Table 1. Bayesian network analysis: conditional probabilities and "what-if" analysis obtained by fixing SNSs usage, active vs. not only uniquely active (mainly passive). Red and blue cells indicates respectively an increase or decrease in scores/percentages after fixing SMU (Social media use) node value.

Variable	BN node name	Category value	Overall	Active	Mainly passive
Social Media Use	SMU	Mainly passive	62.84	0	100
		Active	37.16	100	0
Gender	Gender	Female	68.33	67.32	68.92
		Male	31.67	32.68	31.08
Relationship with face image reproduced in photos or videos	Relat_digitalself	Negative	34.23	29.8	36.85
		Neutral	39.68	41.11	38.83
		Positive	26.1	29.09	24.33
Selfie with friends	Selfie_friends	Never/Rarely	31.53	26.81	34.32
		Sometimes	43.72	42.48	44.45
		Often/Always	24.75	30.71	21.23
Selfie (traditional)	Selfie	Never/Rarely	55.13	51.21	57.45
		Sometimes	29.35	31.56	28.05
		Often/Always	15.52	17.23	14.51
Mirror Selfie	Mirror_selfie	Never/Rarely	59.54	46.17	67.45
		Sometimes	23.28	19.8	25.34
		Often/Always	17.18	34.03	7.21
Age download first social media	Age_1SNS	Early (before 13ys)	20.29	20.25	20.32
		Standard	79.71	79.75	79.68
Selfies taken before getting the one to share on social media	Selfie_number	Don't share	25.75	21.65	28.17
		Satisfied with the first one	7.87	6.78	8.51
		From 2 to 5	34.6	33.99	34.97
		From 5 to 10	22.56	26.16	20.44
		More than 10	9.22	11.42	7.91
Important receiving likes	Like_importance	Don't share	21.04	10.01	27.55
		Not at all-Slightly important	49.69	47.86	50.77
		Quite important-Very important	29.27	42.13	21.67
Used to edit selfies (e.g., adjusting brightness, using filters, etc.)	Editing	No	67.71	66.09	68.66
		Yes	32.29	33.91	31.34
In face-to-face interactions, how much attention do you pay to the other person's "gaze"?	facetoface_you_others	Not at all-Somewhat	8.78	8.64	8.87
		Rather	35.68	35.45	35.82
		Quite a lot-Very much	55.54	55.91	55.32
In digital interactions, how much attention do you pay to the other person's "gaze"?	online_you_others	Not at all-Somewhat	45.84	44.29	46.75
		Rather	37.04	37.94	36.52
		Quite a lot-Very much	17.12	17.77	16.73
In face-to-face interactions, how much do you think others focus on your facial appearance?	facetoface_others_you	Not at all-Somewhat	15.51	15.2	15.69
		Rather	49.23	47.51	50.24
		Quite a lot-Very much	35.27	37.29	34.07
In digital interactions, how much do you think others focus on your facial appearance?	online_others_you	Not at all-Somewhat	33.22	28.61	35.95
		Rather	40.54	42.1	39.62
		Quite a lot-Very much	26.24	29.29	24.43
Appearance Anxiety	total_AAI	[0,4]	28.07	25.83	29.4
		(4,8]	27.64	27.13	27.95
		(8,12]	23.34	24.05	22.91
		(12,31]	20.95	22.99	19.74
FACE-Q Psychological function	FACEQ_psych	[0,50]	28.25	26.58	29.23
		(50,58]	24.44	23.53	24.98
		(58,71]	23.93	24.76	23.44
		(71,100]	23.38	25.13	22.35
FACE-Q social function	FACEQ_social	[0,46]	27.71	26.6	28.36
		(46,55]	24.46	24.27	24.57
		(55,66]	26.09	26.65	25.77
		(66,100]	21.75	22.49	21.31

editing selfies, 61.7% report to never/rarely being worried about conforming to a standard model when using editing tools. Moving to the psychological and aspects related to facial appearance as measured by FACE-Q scales (range 0–100,

with higher scores reflecting a better functioning), median score on psychological functioning was 58.00 (IQR:[50.00, 71.00]), while that on social functioning was 55.00 (IQR:[46.00, 66.00]), with females displaying significantly lower scores than males in both scales. Median score on appearance anxiety scale was 8.00 (IQR:[4.00, 12.00], scale ranges from 0–40), with females displaying significantly higher scores than males.

4 BN Results

The Bayesian network has been estimated using the greedy search algorithm with AIC score function. All quantitative covariates have been categorized using cut-offs known from the literature whenever available or based on statistical criteria (e.g., using medians, quartiles). For categorical variables and Likert items, categories with few values were aggregated. As "blacklists" directions, we set that age, gender, provenance, occupational status, digital skills, age at first SNSs download, cannot be influenced by all the other collected variables related to different SNSs usage forms, selfie behaviours and psychological constructs. Estimated BN is shown in Fig. 1. Once the network has been estimated, propagation analysis has been performed within a "what-if" scenarios analysis. Results are reported in Table 1. Being active in SNSs is associated with an increased selfie taking, not only with friends, but also traditional selfies and mirror selfies, suggesting an increased control of digital self-image. Moreover, appearance anxiety scores are increased, indicating higher appearance concerns, interestingly accompanied by a better psychological well-being in terms of facial appearance as measured by FACE-Q psychological function scale. When interpreting Ego-Alter dialectic within the context of real or digital interactions, when fixing



Fig. 1. Estimated BN showing the dependencies among SNSs use, selfie taking/sharing/editing behaviors, the relationship with digital self-image and facial appearance, internalizing symptoms and other psychological constructs.

online activity, changes in the conditional distributions occur mainly for statements related to digital interactions, with others focusing on the own facial appearance to a larger extent.

5 Conclusions

The study represents a first attempt to jointly evaluate the complex interrelationship among self-image representation in digital context, SNSs usage, psychological wellbeing and selfie behaviour, by adopting a multivariate data-driven technique. Bayesian networks are appealing statistical methods allowing to explore how the studied constructs and facets interact with each other, representing an effective tool to provide evidence and support decisions regarding SNS usage regulations. The data are preliminary and pertain to an ongoing project. The phenomenon under analysis is extremely dynamic due to the inherent nature of SNSs, which constantly adapt to changing user preferences and behaviours, as well as technological advancements in general. Furthermore, behaviours and forms of SNS usage exhibit significant heterogeneity: hence it is a complex phenomenon that cannot be reduced to simplistic explanations, thereby introducing new challenges to be addressed.

The project is ongoing and aims at comparing estimated BNs for different age groups, to identify cross-sectional patterns of online behaviours, psychological profiles and relationship with digital self-image.

References

1. Gioia, F., McLean, S.A., Rega, V., Boursier, V.: What about young adults? Photo manipulation activity? The predictive role of body shame and the mediating effect of body esteem. Curr. Psychol. **43**, 10418–10433 (2024)
2. Cowell, R.G., Dawid, P., Lauritzen, S.L. and Spiegelhalter, D.J.: Probabilistic networks and expert systems: exact computational methods for Bayesian networks. Springer Science & Business Media (2007)
3. Mecatti, F., Vicard, P., Musella, F., Giammei, L.: Bayesian networks versus gender bias. Significance **19**, 16–20 (2022)
4. Musella, F., Vicard, P., De Angelis, M.C.: A Bayesian network model for supporting school managers decisions in the pandemic era. Soc. Indic. Res. **163**, 1445–1465 (2022)
5. Briganti, G., Decety, J., Scutari, M., McNally, R.J. and Linkowski, P.: Using Bayesian networks to investigate psychological constructs: the case of empathy. Psychol. Reports. (2022). https://doi.org/10.1177/00332941221146711
6. Cippà, P.E., et al.: A data-driven approach to identify risk profiles and protective drugs in COVID-19. Proc. Natl. Acad. Sci. **118**, e2016877118 (2021)

Greening Obligations, Avoidance Behavior, and Farms' Environmental and Economic Performance: Insights from FADN

Zelda Brutti, Marzia Freo(✉), and Laura Serlenga

European Commission, Joint Research Centre (JRC), Ispra, Italy
marzia.freo@ec.europa.eu

Abstract. Starting in 2014, a new European Union-wide environmental regulation aimed at curbing soil degradation and promoting biodiversity has required farmers to diversify the type of crops cultivated on their land. A significant part of Direct Payments to farmers were made conditional on compliance with Greening requirements. In this article, we estimate the causal impact of the crop diversification Greening requirement among farms. Leveraging data from the FADN survey, we examine the effects on avoidance behavior, as well as the environmental and economic performance of farms that were not previously compliant with regulatory requirements. We show that the crop diversification measure is effective in increasing the average number of crops cultivated on arable land held by holdings above 10 ha, but also induced avoidance behavior among a portion of farms under the scope of the policy. Additionally, we identify increases in pollutant inputs per hectare. Furthermore, we observe positive shifts in total farm output consistent with some ex-ante simulations.

Keywords: Policy evaluation · Green Direct Payments

1 The Greening Policy

Adapting to a world increasingly concerned about climate change, last cycles of the Common Agricultural Policy (CAP) have been characterized by a progressive increase in the attention to the environment; measures taken to this end included the gradual introduction of environmental criteria to be met by farmers for the allocation of direct aid. Starting with the CAP policy cycle that begun in 2014 (Regulation EU No 1307/2013), so-called 'Greening' payments were introduced as a large, compulsory component of Direct Payments to farmers. This was enforced without altering the overall direct payment budget, so that funds were set aside from basic direct payments and earmarked for the Greening scheme [1].

Greening payments are granted at the farm-level and are meant to support farmers who adopt or maintain farming practices that contribute to the EU's environmental and climate goals. Greening components represent about 23% of the whole CAP budget, or about EUR 12 billion per year. For the individual farmer, Greening payments represent

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 291–296, 2025.
https://doi.org/10.1007/978-3-031-64431-3_49

on average 30% of all direct payments received; non-compliance with the environmental requirements leads to severe financial loss in terms of direct aid, up to 1.25 times the amount of Greening payments, and additional penalties may be imposed by national governments since 2017 [2] and [3].

In this article, we focus on the first Greening requirement, crop diversification (CD), which was a complete novelty of the 2014–2021 CAP cycle and was applied uniformly across the EU. The CD requirement varies according to the size of the farm's arable land (AL); the main conditions of obligation to the CD practice, set out by Article 44 of the Regulation (EU) No 1307/2013, require a) farms from 10 to 30 hectares of AL to cultivate at least 2 crops and dedicate less than 75% of AL to the main crop, and b) farms over 30 hectares of AL to cultivate at least 3 crops and dedicate less than 75% of AL to the main crop and less than 95% to the two main crops [4].

The Greening policies introduced to EU farming have been the subject of a number of evaluation studies over recent years, with a wide variation in terms of geographical scope and methodologies employed. The most frequently pursued objective of the existing evaluations has been quantifying the reaction of EU farmers to the new policy, in order to allow a subsequent assessment of impacts on environmental or economic outcomes. One wave of assessment studies was carried out in very early stages of the 2014–2020 CAP policy cycle, either before its implementation or in its first years, employing mostly predictive methodologies. Exiting literature also features a smaller number of ex-post evaluation studies of the Greening policy, based on actually observed data on farm behavior and changes thereof. The number of ex-post evaluations is considerably smaller than those employing modeling techniques, and a large majority focuses on specific geographic areas, such as subnational regions [6, 7], and [8]. This paper leverages the Farm Accountancy data Network data source to explore two main aspects: 1) the responses of farmers to CD regulation, and 2) the effects of this regulation on farmers' decisions regarding land allocation, usage of pollutant inputs, and farms' performance. We present evidence on avoidance behavior by European Union farmers as a response to the Greening policy and estimate the causal effect of the CD on farms' behaviors and performance.

2 Evaluating the Impact of the Greening Policy Using FADN

The Farm Accountancy Data Network (FADN) data source is employed in this study. FADN is a survey carried out each year and harmonised at the EU level regularly, with a sample of 80 000 farms across Member States (MS) from a population of 10.8 million in 2013. It collects data from commercial farms above a minimum size, and represents 50% of the EU farming population, more than 90% of the production, and the area.

Farmers under the obligation to adhere to the CD had to adhere to the requirements from 2015. This provides a time variation in the treatment which clearly provides the basis for a Differences-in-Differences (DID) causal identification strategy in the relevant population of interest provided that a control group can be found. A consideration is requested for adherence to greening requirements. Some farms already either (or do not adhere) to standard Greening practices prior to the application of the policy. This certainly applies to the Crop Diversification measure where the empirical literature identifies a

significant adherence level of farms with the crop diversification requirements, despite the policy implementation. Farms adherence to CD requirements can be assessed by exploiting the longitudinal nature of FADN, and selecting the panel of farms observed both before and after the policy's implementation, specifically from 2014 to 2018. The treated group may be formed by non-already adherent to obligation conditions farms before 2015, while the control group is formed by the already adherent ones. In this setting the interest is on how outcomes of non-already adherent farms changed conditional to CD implementation. Analyses focusing the pre-policy period reveal significant levels of heterogeneity in crop diversity across countries, farm types and sizes, types, and other structural variables. Consequently, we enhance a Propensity Score Matching (PSM) DID approach. The PSM-DID counterfactual method appears suitable for addressing such extensive heterogeneity, and the absence of a sufficiently long period to observe common trends.

The estimation of the policy impact proceeds as follows. Firstly, a propensity score model is estimated by adopting the Covariate Balancing Propensity Score method. As for the choice of matching variables, we considered: country; type of farm; sizes in hectares of arable land, vineyards, other permanent crops, permanent grassland; livestock units; and paid labour input in Annual Working Unit. Secondly, control units are matched to treated ones based on the nearest neighbor values of the estimated propensity score, after imposing the exact matching condition on country and type of farm. The process is repeated until the balancing and overlap conditions are validated. Thirdly, the impact on each outcome is estimated using a DID specification, by a regression weighted by population weights.

We conducted estimations on the panel of all farms consistently present and under the scope of the CD regulation from 2014 to 2018 for evaluating the policy's impact.

Table 1 displays the Average Treatment Effects on Treated farms, highlighting the differential changes in outcomes observed in 2017, the third year after the introduction of the CD regulation, compared to what would have occurred in its absence. The results indicate a trend of avoidance behavior among farms. Specifically, farms that were not compliant in 2014 and had Agricultural Land (AL) exceeding 30 hectares tended to downgrade to a lower AL classification with less stringent requirements. In detail, farms with 30–50 hectares of AL showed proportions that were 4.5% higher than their counterparts in the counterfactual scenario without the regulation. Similarly, farms with 50–100 hectares of AL exhibited proportions that were 2.4% higher compared to their counterparts.

Significant and consistent differences were observed in the responses to policy requirements when compared to the control group, aligning with the expected direction of the effect. On average, the regulation led to an increase of 0.819 in the number of crops grown by farms. Additionally, there were notable reductions in the shares of arable land allocated to specific crops: a decrease of 19.3% points for the main crop and a 9.8% point decrease for the two main crops combined.

It's also worth noting that the proportion of farms compliant with the regulation reached 71.5%. This suggests a substantial level of adherence to the policy among the farming community, indicating the effectiveness of the regulatory measures in influencing farming practices and land allocation decisions.

The quantitative evidence indicates that there were no significant changes in "Agricultural land" for the treated farms following the policy introduction. However, there was a noticeable redistribution of land among major crop classes. Post-policy, there was a trend of reallocating land from categories like "Cereals" and "Harvested green" to "Dry pulses", "Root and tubers", and "Industrial crops". This shift suggests a strategic adjustment in crop choices by farmers in response to the policy.

Furthermore, there were significant increases in pollutant inputs, as evidenced by total expenditures on fertilisers and pesticides per hectare of utilized agricultural area.

On a positive note, there were shifts in total farm output that aligned well with some pre-policy simulations.

This suggests that while there were challenges and adjustments in response to the policy, farms were able to maintain or even improve their overall production levels. These findings underscore the complex interplay of factors at play when policies are implemented in the agricultural sector, with farms adapting in various ways to meet both regulatory requirements and economic considerations.

Table 1. Estimated ATET on listed outcomes by size of arable land (ha).

	All		of whom:			
			10–30 ha		30–50 ha	
Prevalence of intra-class movements	0.013		0.070	*	−0.045	*
N. of crops	0.819	*	0.715	*	0.852	*
Share of main crop	−0.193	*	−0.243	*	−0.165	*
Share of the two main crops	−0.098	*	−0.060	*	−0.120	*
Share satisfying requirement	0.715	*	0.692	*	0.717	*
Land in hectares allocated to						
Arable land	0.215		1.550	*	−1.181	
Cereals	−1.398	*	0.068		−2.264	*
Dry pulses	0.799	*	0.665	*	0.874	
Root and tubers	0.219	*	0.235	*	0.240	*
Industrial crops	0.976	*	0.147		0.685	
Vegetables and flowers	−0.098		−0.042		−0.032	

(*continued*)

Table 1. (*continued*)

	All		of whom:			
			10–30 ha		30–50 ha	
Fodder and fallows	−0.164		0.526	*	−0.640	
Harvested green and other arable	−0.119	*	−0.047		−0.045	
Other outcomes (value in euro)						
Pesticides per ha	5.6	*	5.6		5.5	
Fertilisers per ha	8.9	*	12.2		5.6	
Total farm output	4844	*	2899		3303	

3 Discussion

The CAP has been one of the most significant policies of the European Union (EU). In 2020, it represented 34.5% of the total EU budget. Despite efforts in recent programming periods to integrate environmental objectives, criticisms have been levied against the effectiveness of current CAP environmental measures in significantly improving environmental and climate performance. The empirical researches on ex-post evaluation of the CAP Greening Policy are often limited to regional or national case studies, and not conclusive. Consequently, there is an urgent call for sustained efforts in conducting high-quality evaluations of the CAP.

This paper present rigorous evidence, at the EU level, based on FADN data about the responses of farmers to CD regulation, and the effects of this regulation on farmers' decisions regarding land allocation. Preliminary results on environmental and economic performance are also discussed.

References

1. European Court of Auditors. Greening: a more complex income support scheme, not yet environmentally effective. Special report No 21. LU: Publications Office (2017)
2. European Union. Regulation No 1307/2013 of the European Parliament and the Council establishing rules for direct payments to farmers under support schemes within the framework o he common agricultural policy and repealing Council Regulation (EC) No 637/2008 and Council Regulation (EC) No 73/2009 (2013)
3. Gocht, A., et al.: EU-wide economic and environmental impacts of CAP greening with high spatial and farm-type detail. J. Agric. Econ. **68**(3), 651–681 (2017)
4. European Parliament (2022). First pillar of the common agricultural policy (CAP): II – Direct payments to farmers. Financing of the CAP (2023)
5. Bertoni, D., et al.: Farmland Use transitions after the CAP greening: a preliminary analysis using markov chains approach. Land Use Policy **79**, 789–800 (2018)

6. Michalek, J.: Environmental and farm impacts of the EU RDP agri-environmental measures: evidence from Slovak regions. Land Use Policy 113 (2022)
7. Kurdyś-Kujawska, A., Strzelecka, A., Zawadzka, D.: The impact of crop diversification on the economic efficiency of small farms in Poland. Agriculture **11**(3) (2021)
8. Imai, K., Ratkovic, M.: Covariate balancing propensity score. J. R. Stat. Soc. Series B Stat. Methodol. **76**(1), 243–263 (2014)



Trimodal Extension Based on the Flexible Generalized Skew-Normal Distribution

Michele Bufalo[1(✉)] and Andrea Nigri[2]

[1] Department of Economics and Finance,University of Bari,Bari, Italy
michele.bufalo@uniba.it
[2] Department of Economics, Management and Territory,University of Foggia, Foggia, Italy
andrea.nigri@unifg.it

Abstract. We propose a novel class of generalized skew-normal densities that improves the flexibility of empirical distributions and can systematically capture skewness, heavy tails, and multimodality. We extend the so-called flexible generalized skew-normal (FGSN) density developed by Y. Ma and M. G. Genton in 2004. The main novelty is the existence of a fifth-order degree term in the polynomial that appears in the cumulative distribution function of such a density. In this case, we prove that our density has at most three modes under certain conditions for the parameters. Leveraging this new approach eases the modeling of data consisting of three subpopulations. For validation, we present examples of both univariate and multivariate cases using demographic data from the Human Mortality Data Base (HMD).

Keywords: flexible generalized skew-normal · multimodality · skew-normal

1 Introduction

The field of skew-normal and associated distribution families has witnessed a significant surge in scholarly attention in recent years. This increased academic focus can be traced back to the foundational work by [1,2], which marked the inception of the univariate skew-normal distribution.
The advancement in skewness regulation within distributions led to the development of the skew elliptical family and other parametric families. For an in-depth exploration of skew distributions, seminal works such as [3]. In particular, Ma and Genton [4] explore the concept of "flexibility" by introducing a generalized (univariate) skew-normal density which admits (at most) two modes; and this is particularly useful for the applications (see, e.g., [5]).
Our aim is to extend the contribution of Ma and Genton, by providing a further flexible generalized skew-normal density which may admits at most three modes. The main advantage of this approach is to capture characteristics such as skewness, heavy tails, and multimodality, in order to enhance the adaptability of empirical distributions, and avoid the use of mixture distributions. Indeed,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 297–302, 2025.
https://doi.org/10.1007/978-3-031-64431-3_50

the proposed distribution requires less parameters, and its analytical tractability allows faster simulations and easier reproducibility from observed data

2 Probabilistic Framework

Let us start by defining our proposed density in the univariate case.

Definition 1. *A random variable X is said to be (standard) flexible generalized skew-normally distributed if its probability density function is given by*

$$f_X(x) = 2\phi(x)\Phi(w_K(x)) \qquad (x \in \mathbb{R}) \tag{1}$$

being ϕ, Φ the (standard) normal probability density and cumulative distribution, respectively, and w_K is an odd polynomial of order K, i.e.

$$w_K(x) = \sum_{i=0}^{\frac{K-1}{2}} a_{2i+1} x^{2i+1} \qquad (K \in \mathbb{N},\ K \text{ odd}). \tag{2}$$

We refer to the density (1) by writing $X \sim \mathcal{FGSN}(0, 1, \boldsymbol{a})$, with $\boldsymbol{a} = (a_1, a_3, ..., a_K)$.
More generally, if $X \sim \mathcal{FGSN}(0, 1, \boldsymbol{a})$ and define the random variable $Y = \zeta + \omega X$, then $Y \sim \mathcal{FGSN}(\zeta, \omega, \boldsymbol{a})$, i.e.

$$f_Y(y) = \frac{2}{\omega}\phi\left(\frac{y-\zeta}{\omega}\right)\Phi\left(w_K\left(\frac{y-\zeta}{\omega}\right)\right) \qquad (y \in \mathbb{R}) \tag{3}$$

When $K = 1$ then the density (1) reduces itself to a skew-normal one. If one chooses $K = 3$ then obtain the density studied by [4]. In particular the authors, proved that when $K = 3$, the density (1) admits at most two modes.
In our work, we consider $K = 5$ and $a_1 = 0$, i.e.,

$$w(x) = a_3 x^3 + a_5 x^5, \tag{4}$$

and give the following result.

Theorem 1. *The density function f_X defined by Eq. (1) with w given by (4), has at most three modes if $a_3 > \varsigma(a_5)$, being*

$$\varsigma(a_5) = -1.5529\,(a_5)^{\frac{1}{2}1_{\{a_5<1\}}} \tag{5}$$

Proof. (Some Hints)
Observe that it is not restrictive assume $a_5 > 0$. Let us start by the (general) computation of $f'_X(x)$

$$f'_X(x) = 2x\phi(x)\eta(x) \tag{6}$$

with

$$\eta(x) = \frac{\phi(w(x))w'(x)}{x} - \Phi(w(x)). \tag{7}$$

In order to prove that f_X has at most three modes we proceed by contradiction. If f_X has more than three modes, then f'_X has at least seven (distinct) zeros. We have to distinguish the following three cases.

i) (Case $a_3 \geq 0$)
It is convenient to write w and w' as

$$\begin{cases} w(x) = xy(a_3 + a_5 y) \\ w'(x) = y(3a_3 + 5a_5 y). \end{cases} \tag{8}$$

In this case, the function η in (7) becomes

$$\eta(x) = \frac{\phi(xy(a_3 + a_5 y))y(3a_3 + 5a_5 y)}{x} - \Phi(xy(a_3 + a_5 y)). \tag{9}$$

As in the case *i*), $f'(x) > 0$ for $x < 0$, since $\eta(x) < 0$. Therefore, it remain to study the case $x \geq 0$. Observe that

$$\eta'(x) = \phi(xy(a_3 + a_5 y))\gamma(y),$$

where

$$\gamma(y) = -y^3(a_3 + a_5 y)(3a_3 + 5a_5 y)^2 + (3a_3 + 5a_5 y) + 10a_5 y - (3a_3 + 5a_5 y)y =$$

$$-25a_5^3 y^6 - 55a_3 a_5^2 y^5 - 39a_3^2 a_5 y^4 - 9a_3^3 y^3 - 5a_5 y^2 + (15a_5 - 3a_3)y + 3a_3. \tag{10}$$

It is easy to verify that $\gamma'' < 0$ for any y, so that γ has at most two zeros in y and two positive zeros in x. Then, η has at most three positive zeros and consequently, f'_X has at most four non-negative zeros.

ii) (Case $\varsigma(a_5) < a_3 < 0$)
First of all, note that $f'(x) < 0$ for any $x \in (0, x^*]$, and $f'(x) > 0$ for any $x \in (-\infty, -x^*]$, being $x^* = \sqrt{-\frac{3a_3}{5a_5}}$. Indeed, one has $(3a_3 + 5a_5 y) > 0$ if and only if $x < -x^* \vee x > x^*$. Hence, it remain to study the case $x \in (-x^*, 0] \cup (x^*, +\infty)$.
Observe that there exists a bijection between $x \in (-x^*, 0) \cup (x^*, +\infty)$ and $\tilde{y} \in (3a_3, +\infty)$, being $\tilde{y} = 3a_3 + 5a_5 x^2$. Under the new variable $\tilde{y}$, the function γ defined in Eq. (10) becomes

$$\gamma(\tilde{y}) = -\left(\frac{\tilde{y} - 3a_3}{5a_5}\right)^3 \left(\frac{\tilde{y} + 2a_3}{5}\right)\tilde{y}^2 + \tilde{y} + 10a_5\left(\frac{\tilde{y} - 3a_3}{5a_5}\right) - \left(\frac{\tilde{y} - 3a_3}{5a_5}\right)\tilde{y} =$$

$$\frac{1}{625a_5^3}\left[-\tilde{y}^6 + 7a_3\tilde{y}^5 - -9a_3^2\tilde{y}^4 - -27a_3^3\tilde{y}^3 + (54a_3^4 - 125a_5^2)\tilde{y}^2 + 375a_5^2(a_3 + 5a_5)\tilde{y} - 3750a_3a_5^3\right]. \tag{11}$$

One can prove that $\gamma(\tilde{y})$ does not have six real roots in $(3a_3, +\infty)$. Therefore η has at most five distinct zeros for any $x \in (-x^*, 0) \cup (x^*, +\infty)$, and consequently, f' has at most six distinct zeros in $(-x^*, 0] \cup (x^*, +\infty)$.

3 Numerical Test

We test our procedure on a real dataset consisting of USA populations gender gaps observed across different ages, during $m = 30$ years, and retrieved from the Human Mortality Database (HMD).

Table 1. p-values from the KS test to assess whether the FGSN distribution is suitable for describing our dataset.

Time series	p-value
Year 1992	0.1217
Year 1993	0.1282
Year 1994	0.1316
Year 1995	0.1995
Year 1996	0.3019
Year 1997	0.2842
Year 1998	0.4086
Year 1999	0.2372
Year 2000	0.4156
Year 2001	0.3581
Year 2002	0.5484
Year 2003	0.3661
Year 2004	0.3535
Year 2005	0.5272
Year 2006	0.4945
Year 2007	0.2678
Year 2008	0.2642
Year 2009	0.3080
Year 2010	0.3109
Year 2011	0.2521
Year 2012	0.3446
Year 2013	0.1861
Year 2014	0.3774
Year 2015	0.2678
Year 2016	0.3800
Year 2017	0.2166
Year 2018	0.1501
Year 2019	0.1482
Year 2020	0.2061
Year 2021	0.1438

Let denote by $\left(x_h^{(j)}\right)_{h\in[1,n]}$the j-th series of log-ratios, for any $j \in [1, m]$. In order to calibrate the parameters

$$\theta^{(j)} = \left(\zeta^{(j)}, \omega^{(j)}, \boldsymbol{a}^{(j)}\right)$$

with $\boldsymbol{a}^{(j)} = (a_3^{(j)}, a_5^{(j)})$, we use the log-likelihood estimation, by solving the following optimization problem, i.e.,

$$\hat{\theta}^{(j)} = \arg\left\{\max_{\theta^{(j)}} \sum_{h=1}^{n} \ln\left(f_X^{(j)}\left(x_h^{(j)}; \theta^{(j)}\right)\right)\right\} \tag{12}$$

where

$$f_X^{(j)}(x_h^{(j)}; \theta^{(j)}) = \frac{2}{\omega^{(j)}}\phi\left(\frac{x_h^{(j)} - \zeta^{(j)}}{\omega^{(j)}}\right)\Phi\left(a_3^{(j)}\left(\frac{x_h^{(j)} - \zeta^{(j)}}{\omega^{(j)}}\right)^3 + a_5^{(j)}\left(\frac{x_h^{(j)} - \zeta^{(j)}}{\omega^{(j)}}\right)^5\right) \tag{13}$$

is the probability density of $x_h^{(j)}$. The cdf of (13) may be obtained by numerical integration of (13), i.e.,

$$F_X\left(x_h^{(j)}\right) = \int_{-\infty}^{x_h^{(j)}} f_X^{(j)}(z; \hat{\theta}^{(j)})\, dz,$$

since no closer formulation is known up now. Figure 1, displays the theoretical CDF (dotted line) vs empirical CDF (solid line) from the fitted FGSN distributions for the first 15 columns of our dataset $\left(x_h^{(j)}\right)_{h\in[1,n]}$. We can see as the simulations fit well the actual data for any age, included the most challenging ones referring to the tails. This is in accordance also with the robustness results obtained from the confidence intervals of the estimated parameters. Similar results occur for the remaining 15 columns of the dataset.

In order to assess the wellness of the FGSN distribution for our dataset, we use the Kolmogorov–Smirnov (KS) test. For generating a sample from our FGSN distribution we consider the well-known piecewise linear distribution, which creates a nonparametric representation of the cumulative distribution function through linear connection between the known cdf values $\left(F_X\left(x_h^{(j)}\right)\right)_{h\in[1,n]}$ and the sample data $\left(x_h^{(j)}\right)_{h\in[1,n]}$. Fixed $j \in [1, m]$, we run the KS test between 10^5 generated FGSN samples (with the same parameters), and then we take the mean of the answered p-values whose results are reported in Table 1. The low p-values obtained from our test confirms the wellness of our approach for any year.

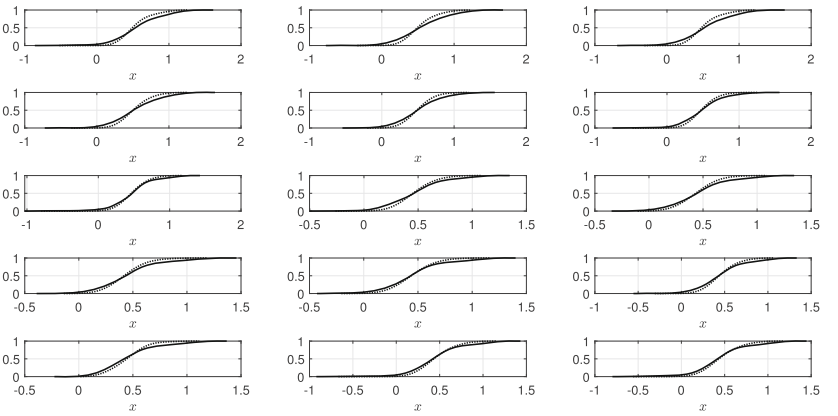


Fig. 1. Theoretical CDF (dotted line) vs empirical CDF (solid line) from the fitted FGSN distributions for the data $\left(x_h^{(j)}\right)_{h\in[1,n]}$ with $j \in [2, m/2]$.

References

1. ArellanoValle, R.B., Branco, M.D., Genton, M.G.: A unified view on skewed distributions arising from selections. Canadian J. Stat. **34**(4), 581–601 (2006)
2. Azzalini, A.: A class of distributions which includes the normal ones. Scandinavian J. Stat., 171–178 (1985)
3. Azzalini. A.: The skew-normal distribution and related multivariate families. Scand. J. Stat. **32**(2), 159–188 (2005)
4. Ma, Y., Genton, M.G.: Flexible class of skew-symmetric distributions. Scand. J. Stat. **31**(3), 459–468 (2004)
5. Mazzuco, S., Scarpa, B.: Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. J. Roy. Stat. Soc. Series A Stat. Soc. **178**(1), 187–203 (2015)

Assessing Literacy Proficiency of Students with a Migratory Background: The Case of Lombardia Region

Alessio Buonomo[1](✉), Stefania Capecchi[1], Francesca Di Iorio[1], and Patrizia Giannantoni[2]

[1] Department of Political Sciences, University of Naples Federico II, Napoli, Italy
{alessio.buonomo,stefania.capecchi,francesca.diiorio}@unina.it
[2] INVALSI, Italian National Institute for Educational Evaluation, Rome, Italy
patrizia.giannantoni@invalsi.it

Abstract. The inclusion of children of immigrants in the destination country schools is a topic of increasing importance in the international literature. In this paper, due to the possibility of using proficiency results in Italian language from the INVALSI (National Institute for the Evaluation of the Education and Training System) standardized tests carried out in 2022–2023, we focus on differences in performance by gender and migration background of students in grade 10 education (corresponding in Italy to the second year of upper secondary school). In this contribution, which is an initial exploratory approach to the topic, we restrain our attention on second-year secondary school students enrolled in Lombardy in the school year 2022–2023. Our analysis takes also into account the different types of secondary schools attended by students and the availability/accessibility of digital tools. We run an ANOVA model considering all possible interactions between the following variables: gender, school type, origin, ownership/availability of digital devices, and language predominantly spoken at home. Our findings indicate statistically different results for first-generation students compared to both natives and second-generation students. Conversely, it can be observed that being a native and having no devices is not significantly different (at 10%) from the condition of being a non-native and having no devices. Thus the results suggests that second generation status actually assimilates the behaviour of the latter to natives.

Keywords: Integration · Migratory background · Literacy · Performances

We acknowledge co-funding from PRIN 2022 entitled The Children of Immigrants Have Grown Up. The transition to adulthood for youth with a migratory background, research unit of the University of Naples Federico II (Protocol 2022PFL7ZB; CUP E53D23010370006). Additional co-funding was secured by grant from University of Naples Federico II, PON "Ricerca e innovazione" 2014–2020 (PON R&I) – AZIONE 4. Title "L'alfabetizzazione digitale come determinante dell'inclusione sociale dei migranti e dei loro figli". CUP E65F21003040003. 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.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 303–308, 2025.
https://doi.org/10.1007/978-3-031-64431-3_51

1 Introduction and Motivation

The inclusion of children of immigrants in the destination country schools is a topic of increasing importance in the international literature [1]. This is even more prominent in the Italian context in which the number of children of immigrants who are born in the country and enrolled in Italian schools is growing remarkably. In fact, when considering the entire Italian school system, they exceed 60% of all foreign students [2, 3]; therefore, examining the different school outcomes by migratory generation should be a prominent tool for designing proper integration policies.

In this paper, due to the possibility of using information on proficiency in Italian language from the INVALSI (the acronym standing for: National Institute for the Evaluation of the Education and Training System) standardized tests carried out in 2022–2023, we focus on differences in performance by gender and migration background of students in grade 10 education (corresponding in Italy to the second year of upper secondary school), as registered on a Rasch scale similar to that used in other well-established international learning surveys, such as OECD-PISA [4–6].

Our analysis takes also into account the different types of schools attended by students and the availability/accessibility of digital tools. Migratory background distinguishes between national and foreign students as classified into first (foreign-born) and second (Italian-born) generations. The interest towards the effect of using digital devices when studying is furthermore relevant due to the Covid-19 pandemic [7], and numerous studies have been proposed on this topic [8]. However, still much remains to be explored, and only a few papers have focused on a comparative analysis by migrant generation.

In this contribution, which is an initial exploratory approach to the topic, we restrain our attention on second-year secondary school students enrolled in Lombardy in the school year 2022–2023, since in this region the largest number of foreign students enrolled is registered (of both first- and second-generation), therefore representing a privileged context for the type of analysis proposed in this manuscript [2].

2 Data, Methods and Results in Brief

Due to the nature of the available variables, which are almost exclusively categorical, it could be convenient to start our preliminary research with an Analysis of Variance in order to disentangle significant interactions. The scope of this study is indeed to add new elements to the already well-known existence of relationships between students' socio-demographic and origin-related variables and their level of school proficiency, especially regarding Italian language knowledge. The point is to better understand and measure such an impact.

The data used in this study are particularly informative regarding both the competencies and skills of the respondents. On the contrary, with regard to information of both economic and socio-demographic backgrounds, more in-depth analyses are difficult to perform.

As mentioned, we focus on the score for proficiency in Italian literacy, a variable on a Rasch scale ranging from 0 to 400 with general average equal to 200. Drivers of interest at this stage of the research are: gender, migratory generation, access to digital tools

(access to, or availability of, home computers, educational software, internet connection, technical manuals), the language predominantly spoken at home, type of secondary school (distinguishing by classical/scientific/linguistic lyceums; other lyceums; technical institutes; professional/vocational institutes). In Lombardy region, during the period under analysis, grade 10 students account for 77,143, equally split by gender (female students are 50.4% of the total). About 10.3% of the respondents have an immigrant background, and among them about 70% are second-generation children. With respect to school type, 36% of the total attend a Lyceum (scientific, classical, linguistic, or other); 34% a technical institute; 13% a professional institute, 17% other high school. Table 1 summarizes the average score by gender and pupils' origin, which, as it can be observed, seems to confirm well-grounded evidence [2, 9, among many others] (Table 2).

Table 1. Average score for proficiency in Italian literacy by gender and origin background, standard errors in brackets

M native	M IG	M IIG	F native	F IG	F IIG
200.9 (0.2)	173.0 (1.0)	183.7 (0.7)	210.9 (0.2)	177.0 (0.9)	192.0 (0.5)

Legend: M = males; F = females; IG: First generation migrants; IIG: Second generation migrants

Table 2. Average score for proficiency in Italian literacy by origin background and possession of digital devices.

Origine	No digital device	At least 1 digital Device	At least 2 digital devices	At least 3 digital devices	4 digital devices	Total
Natives	160.05	192.70	207.87	211.68	211.73	210.55
I gen. migrants	153.58	155.94	173.22	180.75	177.29	175.82
II gen. migrants	150.90	176.22	188.23	191.42	188.22	188.60
Total	156.73	184.50	202.39	208.05	208.12	206.29

About 43% of both first- and second-generation migrant students are enrolled in a technical school, whereas about 23% are enrolled in a classical/scientific/linguistic high school. When looking at natives, the percentages are clearly reversed: 39% of nationals are enrolled in a Lyceum, while 32% attend a technical institute.

Non-native males enrolled in technical institutes reach 57%, while the females are nearly 32%; at the (classical/scientific/linguistic) Lyceums, male students with a migratory background account for 20%, while females are 27%; in other curricula (in particular, in the so-called Humanities Lyceum), non-native female students reach 15%, while males 4.5%. The remaining share is distributed among professional institutes. Notice that among natives, the proportion of those attending a professional school is about 11%,

and it is quite the same when distinguishing by gender. Among students with a migratory background, this proportion comes to 28% of first-generation female students while 20% are first-generation males. As for the second generation, 18% of females *vs.* 15% of males are enrolled in professional schools.

These aspects alone related to the segregating choices of school types would merit further study, which we hope to conduct afterwards, also in light of the effects of digitization on school learning, in a longitudinal perspective.

Due to a sizable proportion of missing data being deleted from the analysis, we examined 61,514 complete questionnaires.

In view of the nature of the data as a first analysis with an essentially exploratory scope, we run an ANOVA model considering all possible interactions between the following variables: gender, school type, origin, ownership/availability of digital devices, and language predominantly spoken at home. The underlying rationale is to take into account both the gender-related preferences towards different types of school and that linked to one's origin, not disregarding the fact that, as widely known in the literature [10], speaking at home, predominantly, a language other than the local one can make the destination language learning more challenging. Thus, our goal, rather than to verify the presence of a previously identified effect, is to quantify it and compare the extent of these effects across natives and non-natives.

In the general model accounting for all the possible interactions, not reported for the sake of space, we observe the significance of the single treatments on the level of proficiency in Italian language, as well as those of the specific interactions as follows: between gender and school type, gender and origin, origin and language spoken at home, availability of digital devices and Internet connection, digital and language, school type and digital. The only significant higher-level interactions are: gender and origin, school type and digital, origin and digital, consistently with the descriptive analyses as reported. We then estimate a constrained model in order to highlight the existence of statistically significant differences between these interactions, the results of which are shown in Table 3.

Focusing on the interaction between origin and access to digital devices, it can be observed that being a native and having no devices is not significantly different (at 10%) from the condition of being a non-native and having no devices. We can therefore imply that the relevant impact on the proficiency in Italian is that of the "digital isolation" rather than having or not a migratory background.

The pattern is different if the availability of at least one device is considered, and in particular when considering the presence of Internet connection at home

In this case, the first-generation students show a statistically different result as compared to both natives and second-generation students, thus suggesting that second generation status actually assimilates the behaviour of the latter to natives. Similar results are obtained with regard to gender: as expected females perform better than males [9] and this can be observed among both natives and second generation, while there seems to be no significant difference by gender in first generation respondents.

These initial results encourage us to design future further investigations. It will certainly be of great interest to extend this analysis to additional regions, also from a

Table 3. Restricted ANOVA model estimates

Source	Partial SS	Df	MS	F
Model	23315277	81	287842.92	281.55
Gender	28875.908	1	28875.908	28.25
Origin	41192.257	2	20596.128	20.15
Gender#Origin	14635.437	2	7317.7185	7.16
School type	606333.36	3	202111.12	197.70
Digital (devices)	138666.87	4	34666.716	33.91
School type#Digital	27310.654	12	2275.8878	2.23
Origin#Digital	14611.579	8	1826.4474	1.79
Language (at home)	83688.712	1	83688.712	81.86
Origin#Language	53935.83	2	26967.915	26.38
Gender#Origin#Digital	39110.259	12	3259.1882	3.19
School type # Origin # Digital	58036.591	30	1934.553	1.89
Digital#Language	10667.575	4	2666.8937	2.61
Residual	62804040	61,432	1022.3343	
Total	86119317	61,513	1400.0182	
Adj R-squared = 0.2698		Root MSE = 31.974		

comparative perspective. In addition, it will be of interest to compare the results considering both pre-Covid-19 and post-pandemic school years in order to assess whether important differences have occurred over time and to evaluate the impact of the pandemic on Italian education, both in terms of performance and in terms of the role-played on digital literacy. Lastly, in addition to considering Italian language, it will be of interest to consider the average scores obtained in mathematics.

References

1. Eurostat: Early leavers from education and training. Statistics Explained. https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Early_leavers_from_education_and_training#Early_leavers_from_education_and_training_.E2.80.93_today_and_a_historical_comparison. Accessed 05 Mar 2024
2. MIUR: Alunni con cittadinanza non italiana, anno scolastico 2020/2021, Direzione generale per i sistemi informativi e la statistica, Ufficio di statistica (2022)
3. Strozza, S.: L'inserimento scolastico dei figli degli immigrati: una questione aperta. La Rivista delle Politiche Sociali **3**(3), 127–146 (2015)
4. INVALSI: Quadro di Riferimento delle prove INVALSI di Italiano. https://invalsi-areaprove.cineca.it/docs/file/QdR_ITALIANO.pdf. Accessed 05 Feb 2024
5. OECD: Programme for International Student Assessment's (PISA) periodic testing program on student performance. https://www.oecd-ilibrary.org/education/pisa-2022-results-volume-i_53f23881-en. Accessed 28 Feb 2024

6. INVALSI: Rapporto Invalsi 2023. Roma (2023)
7. UNESCO: COVID-19 Educational Disruption and Response. https://en.unesco.org/covid19/educationresponse. Accessed 27 Feb 2024
8. Baker, J.: The kids who will never return to school after COVID-19. Sydney Morning Herald. https://www.smh.com.au/national/the-kids-who-will-never-return-to-school-after-covid-19-20200411-p54j0e.html. Accessed 05 Mar 2024
9. Strozza, S., Buonomo, A., Muccitelli, P., Spizzichino, D., Casacchia, O., Gabrielli, G.: I giovani stranieri e la scuola. In: Perez M. (Ed.), Vita e percorsi di integrazione degli immigrati in Italia. Istituto Nazionale di Statistica, pp. 193–215, Roma (2018)
10. Bonifazi, C., Buonomo, A., Paparusso, A., Strozza, S., Vitiello, M.: La conoscenza dell'italiano e i processi di integrazione. In: Cadeddu, M.E., Marras, C. (eds.), Linguaggi, ricerca, comunicazione. Plurilinguismo e migrazioni, Focus del Consiglio Nazionale delle Ricerche (CNR), pp. 97–114, Roma (2019)

Modeling and Synthesis of GPS Trajectories for Privacy-Preserving Mobility Data Generation

Arianna Burzacchi[1(✉)], Lise Bellanger[2], Klervi Le Gall[2], Aymeric Stamm[2], and Simone Vantini[1]

[1] MOX Laboratory, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 2, 20133 Milano, Italy
arianna.burzacchi@polimi.it

[2] Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, 2 Rue de la Houssinière BP, 92208 Nantes, France

Abstract. This research introduces a novel methodology for generating synthetic mobility data from GPS trajectory records, addressing privacy concerns while maintaining data utility. The approach employs Functional Data Analysis (FDA) to model trajectories as multidimensional curves, capturing variations in latitude, longitude, and time-to-end. Starting from a dataset of smartphone GPS records, through spline interpolation and a two-step synthesis procedure, synthetic trajectory functions are generated, one for each original function, considering the instances with the shortest amplitude and phase distance from the reference function, performing a weighted average of their Square-Root Velocity Functions (SRVFs), and then mapping back to the original functional space. Offering scalability and adaptability to various functional inputs, this methodology advances privacy-preserving data synthesis, opening new directions for mobility data analysis.

Keywords: Synthetic Data Generation · GPS Data · Functional Data Analysis · Phase and Amplitude distance · Square-Root Velocity Function

1 Introduction

In the last decade, the issue of privacy has become central when discussing data collection, manipulation, and publishing. Both governmental entities and companies have focused their attention on the value of data and the importance of protecting the users to whom such data refers. Traditionally, privacy protection involves anonymizing data by censoring identifiers and other sensitive information, but this technique has been demonstrated to be outdated. With a constantly increasing number of observations collected across various domains, each individual in any given database becomes increasingly distinct from others

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 309–314, 2025.
https://doi.org/10.1007/978-3-031-64431-3_52

and thus easily re-identifiable. In her pioneering work, Latanya Sweeney demonstrated that the simple combination of zip code, date of birth, and gender is unique for 87% of the United States population and proves to be sufficient for re-identifying individuals [1]. Focusing on human mobility, subsequent studies have confirmed this insight by demonstrating that privacy is not fully preserved by removing personal data, as individuals can be re-identified based on the uniqueness of their trajectories [2]. For this reason, traditional techniques for anonymizing datasets have been questioned in favor of innovative methodologies overcoming this limitation.

The generation of synthetic data has emerged as a solution to the problem of re-identification. Synthetic data is defined as artificial data produced from original data and a generative model trained to reproduce the characteristics and structure of the original data. Therefore, synthetic data is not associated with any particular individual. Instead, it captures the overall behavior of the original dataset while preserving its characteristics for analysis and its utility in applications.

Both companies and research institutions have worked to create synthetic data using a privacy-compliant and utility-preserving approach. Classic examples include Synthetic Data Vault (SDV) [3] and Gretel [4]. Regarding the generation of synthetic mobility data utilizing GPS trajectory data as original data, algorithms can be found both in the field of artificial intelligence [5–7] and in that of statistics and probability [8–10].

Fitting in this context, the work presented in this manuscript proposes the construction of a new methodology for generating synthetic mobility data, specifically taking GPS trajectory data as input. From a modeling perspective, we framed the analysis within the domain of Functional Data Analysis (FDA) [11,12]. Each trajectory is indeed modeled as a multidimensional curve that describes the variation of latitude, longitude, and time-to-end as a function of the record ordering in the trajectory sequence. Subsequently, attention is given to the choice of the best mathematical embedding for the current analysis, questioning the geometrical invariance properties of the curves to define a representative functional distance for the specific application of interest. Based on this modeling, virtual curves are constructed from the characteristics of the original curves present in the k-nearest neighborhood of each curve.

This method stands out for its modeling innovation, as no previous study has treated trajectories as curves within an appropriate mathematical space. This perspective opens up new possibilities for the analysis and interpretation of mobility data. Additionally, the proposed data synthesis procedure for trajectories is highly scalable, allowing its application to a wide range of functional inputs, making it flexible and adaptable to various situations and application contexts, even beyond the specific field of human mobility.

2 Outline of the Methodology

The proposed algorithm utilizes an input set of GPS measurements describing the trajectories of distinct anonymized users. This raw data is provided as GPS

signals $P = (c_1, c_2, t)$, representing the record measured at time t at coordinates c_1 and c_2, grouped in N trajectories. Each trajectory is hence a sequence of points ordered with respect to the timestamp, $T_j = \{P_1, \ldots, P_{n_j}\} \quad \forall j = 1, \ldots, N$.

Framing the analysis in the FDA domain, we model the sequences of points as multidimensional functions. Specifically, the domain of each component is the interval $I = [0, 1]$, with $x \in I$ representing the temporal percentage of completion of the trajectory from its beginning to its end. The three dimensions of the functions represent the variation of coordinates (f^1, f^2) and the time elapsed during the trajectory measurements (f^3), fully capturing the information of the GPS data. Each trajectory T_j can be hence seen as the function $\boldsymbol{f}_j$:

$$\begin{aligned} \boldsymbol{f}_j &: I \to \mathbb{R}^2 \times \mathbb{R}_+ \\ \boldsymbol{f}_j &: x \mapsto [f_j^1(x), f_j^2(x), f_j^3(x)] \end{aligned} \tag{1}$$

To move from the discrete sequences of points T to their functional representation $\boldsymbol{f}$, the smoothing method via splines is employed for each component.

Once the functions are constructed, they are provided as input to the data synthesis method. From a dataset of N functions $\{\boldsymbol{f}_j\}$, an equally-sized synthetic dataset of functions $\{\boldsymbol{f}_j^{(s)}\}$ is generated using the following proposed method. The synthesis is performed in a two-step procedure summarized in the following points:

1. Computation of similarities between all the N functions of input
2. For each function $j = 1, \ldots, N$, search of the K nearest functions and computation of a suited weighted average of these as the new synthetic function

For the first step of the algorithm, it is necessary to define the notion of distance between the functions under analysis. We opt for the decomposition of distance into the two dimensions of amplitude distance and phase distance. Indeed, from the applicative point of view, the separation of amplitude and phase components allows us to account for the variability in the sampling frequency of GPS data that may cause phase misalignment. Distances are computed following the approach in [13,14], exploiting the definition of the Square-Root-Velocity Function (SRVF) and the Fisher-Rao Metric. The choice arises for the flexibility of this modeling framework to work with equivalence classes, and not just individual functions. A detailed discussion of the metric can be found in the reference works [13,14]. Here follows a description of the basic concepts to gain a sufficient understanding of the proposed methodology.

Let $\mathcal{F}(I) = \{\boldsymbol{f} : I \to R^2 \times R_+ \text{ absolutely continuous}\}$ be the space of the multidimensional input functions. Given any function $\boldsymbol{f} \in \mathcal{F}(I)$, its SRVF $\boldsymbol{q} \in L^2(I)$ is defined as:

$$\boldsymbol{q}(x) = \frac{\dot{\boldsymbol{f}}(x)}{\sqrt{\|\dot{\boldsymbol{f}}(x)\|}} \quad \forall x \in I \tag{2}$$

Define $\Gamma = \{\gamma : I \to I \text{ warping function}\}$ as the set of all warping functions on I, i.e., the set of non-decreasing monotonic diffeomorphisms on I. Due to this

definition of Γ, the SRVF of a warping function is $\psi = \sqrt{\dot{\gamma}}$. It is to be noted that the SRVF of any warping lies in the L_2 sphere, i.e., $\psi \in S_1^{\infty}(I) = \{g \in L^2(I) \mid \|g\|_{L_2(I)} = 1\}$.

The amplitude distance between two functions $\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathcal{F}(I)$ is defined as the minimum L_2 distance between aligned SRVFs, that is:

$$D_a(\boldsymbol{f}_i, \boldsymbol{f}_j) = \min_{\gamma \in \Gamma} \| \boldsymbol{q}_i - (\boldsymbol{q}_j \circ \gamma)\sqrt{\dot{\gamma}} \|_{L_2(I)} \tag{3}$$

The amplitude distance is hence the Fisher-Rao distance between the two functions. Given $\gamma_{opt} \in \Gamma$ as the minimizer of the quantity in Eq. 3 and $\psi_{opt} \in S_1^{\infty}(I)$ as its SRVF, given the identity warping and its SRVF $\psi_0 \in S_1^{\infty}(I)$, i.e., the constant function $\psi_0 : x \mapsto 1$, there can be defined the phase distance as the distance between ψ_{opt} and ψ_0 computed on the sphere $S_1^{\infty}(I)$, that is:

$$D_p(\boldsymbol{f}_i, \boldsymbol{f}_j) = \| \psi_0 - \psi_{opt} \|_{S_1^{\infty}(I)} = \cos^{-1}(< \psi_0, \psi_{opt} >) \tag{4}$$

Given this analysis setting, in the first phase of the synthesis method, amplitude and phase distances are computed for all pairs of functions following the definitions in Eq. 3 and Eq. 4. Then, the final distance is calculated as a convex combination of the two, as in Eq. 5. The optimal weight $\delta \in [0, 1]$ can either be chosen based on prior knowledge, or exploiting techniques of source matching [15].

$$D_{tot}(\boldsymbol{f}_i, \boldsymbol{f}_j) = \delta \, D_a(\boldsymbol{f}_i, \boldsymbol{f}_j) + (1 - \delta) \, D_p(\boldsymbol{f}_i, \boldsymbol{f}_j) \tag{5}$$

The second step of the synthesis method is repeated for all the functions of the input dataset. Given a reference function, its K closest functions with respect to the distance D_{tot} are found, and their SRVFs are used to compute an average function weighted on the inverse of the relative distances. Noise is added to the resulting function to add variability, enhancing the privacy compliance of the approach. Once mapped back to the original functional space through the inverse SRVF transformation, the collection of all such mean functions constitutes the output synthetic dataset.

The process recalls the k-nearest neighbors (k-nn) algorithm due to the search of the most similar instances. However, here the weighted averaging process is made stochastic by sampling weights from a Dirichlet distribution, thereby enhancing the privacy protection of the users in the original dataset, diverging from the usual k-nn. Specifically, denoting $\boldsymbol{d} = \{d_k\}$ as the K distances between the reference function and the K nearest functions, firstly there is defined the vector of parameters $\boldsymbol{\alpha}$ of elements $\alpha_k = \frac{\alpha_0 g(d_k)}{\sum_{n=1}^{K} g(d_n)}$, and then the vector of weights $\boldsymbol{w} = [w_k]$ is randomly sampled as: $\boldsymbol{w} \sim Dirichlet(\boldsymbol{\alpha})$. For the choice of the parameters K and α_0, the distribution of the distances between synthetic and original functions is considered, looking for both local maxima in the minimal distances and local minima in the maximal distances. The function $g : \mathbb{R}^+ \to \mathbb{R}^+$ can be any monotone non-increasing function, such as $g(x) = \frac{1}{x+1}$ or $g(x) = e^{-x}$, describing an inverse proportionality between the weights and the distance.

Finally, the methodology is validated by evaluating two distinct aspects: the similarity and the privacy compliance of the synthetic dataset with respect to the original one. Indeed, any synthesis method is expected to generate data that are similar to the original ones, with comparable summary statistics and general features, so that the overall information contained in the original dataset is preserved in the synthetic one. At the same time, it is necessary to avoid making individual user information available preserving their privacy. Hence, the method is evaluated considering privacy metrics such as k-anonymity.

3 Applications and Main Results

The proposed algorithm is applied to the case study of Milan, Italy, in the period of interest between December 1st and December 20th, 2019. The dataset of GPS traces has been supplied by Cuebiq Inc., a company specializing in location intelligence, through their Data for Good initiative. Data has been firstly filtered, selecting non-noisy data occurring in daily hours on weekdays, and belonging to trajectories that end in the same Point of Interest (POI), that is, the Bovisa campus of Politecnico di Milano and surroundings, in the north of Milan. The dataset is composed of 11 928 records of 1 216 distinct trajectories collected from 745 smartphone users.

The proposed methodology is proven to generate valuable synthetic data. In terms of privacy, the synthetic dataset satisfies the expected standards. Moreover, the visual inspection of some representative trajectories highlights the similarity between the two datasets. The graphical visualizations of heatmaps and

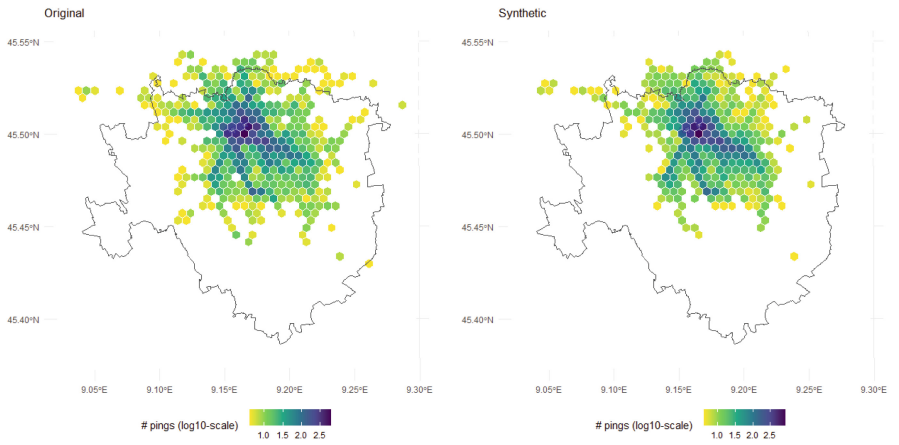


Fig. 1. Heatmaps showing the number of trajectories per hexagon in the study area for both original data (left) and synthetic data (right). Synthetic curves are computed using $K = 5$, $\alpha_0 = 3$, and $g(x) = e^{-x}$; then records are sampled from curves using sampling rates comparable to the original ones.

the trajectory distributions demonstrate that both the original dataset and the synthetic one contain consistent information. An example of heatmap for both original and synthetic data is shown in Fig. 1.

References

1. Sweeney, L.: Simple demographics often identify people uniquely (2000). http://dataprivacylab.org/projects/identifiability/
2. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. Scientific Reports **3**(1), 1376 (2013). https://doi.org/10.1038/srep01376
3. Datacebo Inc. Synthetic data vault. https://docs.sdv.dev/sdv. Accessed 20 Sep 2023
4. Gretel Labs. Gretel: The synthetic data platform for developers. https://gretel.ai/. Accessed 20 Sep 2023
5. Blanco-Justicia, A., Jebreel, N.M., Manjón, J.A., Domingo-Ferrer, J.: Generation of synthetic trajectory microdata from language models. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases, pp. 172–187. Springer International Publishing, Cham (2022)
6. Wang, H., Gao, C., Wu, Y., Jin, D., Yao, L., Li, Y.: PateGail: a privacy-preserving mobility trajectory generator with imitation learning. In: Proceedings of the AAAI Conference on. Artif. Intell. **37**(12), 14539 (2023). https://doi.org/10.1609/aaai.v37i12.26700
7. Zhang, J., Huang, Q., Huang, Y., Ding, Q., Tsai, P.W.: DP-TrajGAN: a privacy-aware trajectory generation model with differential privacy. Futur. Gener. Comput. Syst. **142**, 25 (2023). https://doi.org/10.1016/j.future.2022.12.027
8. Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., González, M.C.: The time Geo modeling framework for urban mobility without travel surveys. Proc. Natl. Acad. Sci. **113**(37), E5370 (2016). https://doi.org/10.1073/pnas.1524261113
9. Bindschaedler, V., Shokri, R., Gunter, C.A.: Plausible deniability for privacy-preserving data synthesis. Proc. VLDB Endow. **10**(5), 481–492 (2017). https://doi.org/10.14778/3055540.3055542
10. Pappalardo, L., Simini, F.: Data-driven generation of spatio-temporal routines in human mobility. Data Min. Knowl. Disc. **32**(3), 787–829 (2017). https://doi.org/10.1007/s10618-017-0548-4
11. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, Springer (2005)
12. Srivastava, A., Klassen, E.P.: Functional and shape data analysis, Springer (2016)
13. Kurtek, S., Srivastava, A., Klassen, E., Ding, Z.: Statistical modeling of curves using shapes and related features. J. Am. Stat. Assoc. **107**(499), 1152 (2012). https://doi.org/10.1080/01621459.2012.699770
14. Tucker, J.D., Wu, W., Srivastava, A.: Generative models for functional data using phase and amplitude separation. Comput. Stat. Data Anal. **61**, 50 (2013). https://doi.org/10.1016/j.csda.2012.12.001
15. Bellanger, L., Coulon, A., Husi, P.: Une méthode de classification ascendante hiérarchique parcompromis: hclustcompro. In: 9e Conférence Internationale Francophone sur la Science des Données (CIFSD) (CIFSD, Mohamed QUAFAFOU, Marseille, France) (2021). https://hal.science/hal-03280918

A Modeling Approach to Investigate Species-Environment relationship in Coral Areas Using *in-Situ* observations

Crescenza Calculli[1(✉)], Diego Battagliese[1], Porzia Maiorano[1,2], Angela Carluccio[1,2], and Gianfranco D'Onghia[1,2]

[1] University of Bari Aldo Moro, Bari 70124, Italy
crescenza.calculli@uniba.it
[2] Consorzio Nazionale Interuniversitario per le Scienze del Mare (CoNISMa), Roma 00196, Italy

Abstract. The distribution of relevant benthopelagic species in two cold-water coral provinces of Central Mediterranean is investigated. Using *in-situ* data from video system surveys, the work explores coral habitat roles and environmental influences on biodiversity. Distributional regression models estimate species abundance, emphasizing the critical importance of selecting appropriate abundance measures due to their direct influence on modeling choices. GLM and GAMLSS frameworks are used to model count and non-negative abundance data, allowing simultaneous evaluation of several area-specific environmental drivers. Results reveal variations in estimated abundances owing to environmental factors, aiding ecologists in assessing species abundance in these ecosystems.

Keywords: species abundance · baited-lander survey · overdispersion · GLM · GAMLSS

1 Introduction

Marine ecosystems currently face several direct threats such as habitat degradation, overexploitation, pollution, climate change, and non-indigenous species introduction, that endanger ecosystem integrity, functionality, and species distribution patterns. In this context, understanding how different environmental factors affect species distribution is recognized as a pivotal subject within environmental and ecological sciences, especially for the implementation of conservation and restoration plans [3]. Ecologists often rely on abundance estimates to evaluate the conservation status of species, considering abundance variation as a crucial aspect of a species' response to habitat quality [2]. However, the intricate relationship between environmental patterns and ecological processes requires innovative research approaches to ensure accurate abundance predictions. Since continuous sampling of a species' population is typically impractical, the analysis of environmental drivers influencing species abundance involves surveying representative subsets of the area or population of interest, fitting models to assess

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 315–320, 2025.
https://doi.org/10.1007/978-3-031-64431-3_53

the relationships between observed data and environmental factors, and using these models to predict abundances in unsampled areas. Despite the wide-spread adoption of this approach, implementing such models entails tackling complex statistical challenges, particularly in the case of mobile organisms with spatio-temporally varying distributions [1]. In this spirit, we propose to analyze *in-situ* observation data for exploring the functional role of corals in benthopelagic fauna diversity and abundance within two cold-water coral (CWC) areas in the central Mediterranean. The data are gathered from surveys that use the Marine Environment Monitoring system (MEMO) baited lander, which offers a low-impact surveying approach for seafloor monitoring compared to traditional methods that could harm the benthic habitat. These surveys provide different measures of coral habitat abundance, leading to different modeling choices that take into account the response distribution assumptions. Distinguishing between count and non-negative abundance responses, we consider Poisson/Negative Binomial and Gamma regression models respectively addressing excess of variability in data as needed [6]. We propose estimating abundances using Generalized Linear Models (GLMs) and Generalized Additive Models for location, scale, and shape (GAMLSSs), which allow for flexible modeling of complex relationships between survey data and environmental covariates [8]. This approach is particularly useful for ecosystems with limited prior knowledge of species-environment relationships, as it enables the simultaneous selection of covariates to explain local variation in species distribution.

2 Study Area and Data Collection

The study area lies within the Central Mediterranean Sea along the Apulian margin, situated between the southern Adriatic Sea and the northern Ionian Sea (see Fig. 1). This region is characterized by several morphological and geological formations that shape the continental shelf. Along the Apulian margin, a belt of cold-water coral communities linked by water masses flowing from the southern Adriatic to the northern Ionian, has been recently identified [4]. The complex morphological structure of the Bari canyon hosts a diverse community of deep-water cnidarians. Benthopelagic fauna in the Bari canyon includes teleost fishes like *Conger conger*, *Helicolenus dactylopterus*, *Merluccius merluccius*, *Phycis blennoides*, and *Pagellus bogaraveo*, with *Galeus melastomus* being the most abundant cartilaginous fish [5]. The SML cold-water coral (CWC) province is located near Cape Santa Maria di Leuca (Italy). Recent mapping activities revealed complex topographical features, including hummocky seafloor sectors with mound-like reliefs and fault-scarps indicating stiff or hardened older sediments [9]. Living coral colonies are found at lower depths with a patchy distribution associated with block patterns and clustered tapping. Previous studies highlighted the importance of the entire SML CWC province as a diverse multi-dimensional habitat, serving as a refuge area for various species, including teleost fishes and providing protection from fishing pressure [5]. Abundance data were obtained from video surveys conducted as part of the Coral-FISH project and

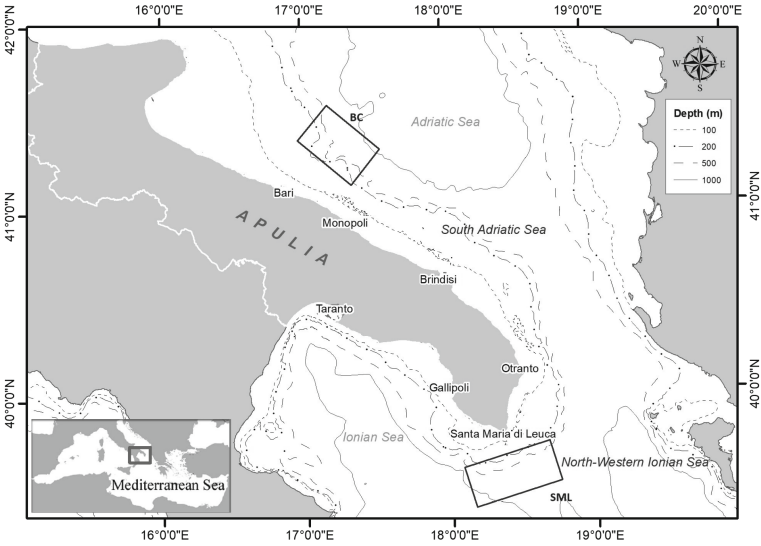


Fig. 1. The study area in the central Mediterranean Sea. Delimited areas indicate Bari Canyon (BC) and Santa Maria di Leuca (SML) CWC provinces.

the Italian OBAMA-PRIN research. For these surveys, the MEMO baited lander equipped with two digital cameras capturing oblique and horizontal views for the coverage of small-scale areas were used. Deployments occurred between June 2010 and July 2019, covering two habitat types: heterogeneous and complex (HC) with live coral colonies such as CWC and muddy bottom (MB) non-coral habitats, mainly characterized by fine-grained sediments. A total of 27 deployments were carried out along the Apulia margin, recording 123 h in HC habitats and 73 h in MB habitats. Species identification was conducted visually, with abundances standardized to video recording time. For both habitats, the number of species (*NSpecies*) and the number of individuals (*Nind*) were obtained. Species abundance was also assessed based on the time of first arrival (1^{st} *arrival time*) after the deployment and the maximum number of individuals observed in a single frame (*MaxN*) per species. Environmental factors including habitat type (HC vs MB), geographical area (Southern Adriatic, SA corresponding to BC province vs Northern Ionian, NI corresponding to SML province), deployment depth (dep), total time of observation in hours (hrec) and the survey season, Spring (SP), Summer (SU) and Autumn (AU), are considered to investigate their effects on local relevant species abundances.

3 Modelling Approach for Abundance Data

For the most relevant species in the two areas (*Conger conger, Helicolenus dactylopterus, Merluccius merluccius, Pagellus bogaraveo, Phycis blennoides, Galeus melastomus*), abundances in terms of *NSpecies*, *MaxN*, *Nind* and 1^{st} *arrival time*

are modeled using regression-like frameworks. Different specifications are evaluated to appropriately capture the dependence between these responses and area-specific environmental factors. For count responses (*NSpecies*, *MaxN*, *Nind*), Poisson and Negative Binomial (NB) regression models are considered. These models are commonly used to describe occurrences and count data where lower values of a variable are more frequent. Conditional on explanatory variables, Poisson regression assumes equality between the mean and variance. However, when the estimated variance exceeds the estimated mean, alternative methods are required to prevent underestimation of standard errors and biased inference (e.g. quasi-Poisson model [10]). Negative Binomial (NB) regression can be used as an effective alternative to the Poisson distribution to deal with overdispersion. The Negative Binomial (NB) model includes an additional term to accommodate the variance excess, resulting in a 2-parameters distribution. Thus, while the random variable $Y \sim \text{Pois}(\mu)$ has $Var_{\text{Pois}}(Y) = \mu$, for $Y \sim \text{NB}(\mu, \kappa)$ variance is $Var_{\text{NB}}(Y) = \mu + \kappa\mu^2 (\mu > 0 \text{ and } \kappa > 0)$ with overdispersion represented by the multiplicative factor $1 + \kappa\mu$, which depends on μ. Moreover, we consider Gamma distribution assumption to model positive right-skewed data concerning the *1st arrival time* response. In this case, the distribution is a function of the shape and scale parameters.

Two general frameworks are proposed (and compared) to model abundances of relevant species. First, Poisson GLMs are estimated to model count abundance data in terms of *NSpecies*, *MaxN* and *Nind*. In this case, the mean of the Poisson distribution is connected to the linear predictor, η, using the log-link as $log(\mu) = \eta = X'\beta$, where β is a vector of regression coefficients associated with covariates. To account for the potential overdispersion in count data, we propose a second framework named Generalized Additive Models for Location, Scale and Shape GAMLSS [8] to estimate NB and Gamma models. This method allows for the modeling of several distributions by explicitly capturing various distributional parameters, including mean, dispersion, skewness, and kurtosis, which can all be linked to covariates. Thus, in the GAMLSS framework for count data, the mean μ and the dispersion σ of an assumed NB distribution are modeled in terms of covariates using a suitable link function: $log(\mu) = \eta_1 = X'\beta_1$ and $log(\sigma) = \eta_2 = X'\beta_2$. The same specification is considered for estimating Gamma model. Maximum likelihood estimates (MLEs) for GLMs are obtained using the Iterated Weighted Least Squares (IWLS) algorithm in the `glm` function of the `stats` package of `R` software [7]. Parameter estimates for GAMLSSs are derived by maximizing a penalized likelihood function using the `gamlss` function of the `gamlss` package of the same software [8].

4 Results

Based on preliminary EDA (not presented for brevity sake), regression-type models are estimated accounting for the response variables characteristics and the potential effects of environmental factors affecting abundances. Considering the *NSpecies* response for which no overdispersion in the data is detected (dispersion

Table 1. Estimated coefficients from fitted regression models for each abundance response considering relevant species. Stars correspond to: ***, p-values$<$ 0.001; **, p-values$<$ 0.01; *, p-values$<$ 0.05. MB stands for Muddy Bottom; NI stands for northern Ionian; SP stands for Spring and SU for summer.

		NSpecies (GLM)	*MaxN* (GAMLSS)	*Nind* (GAMLSS)	*1st arrival time* (GAMLSS)
$log(\mu)$	Intercept	0.251	4.980^{***}	5.448^{***}	0.689
	$habitat_{[MB]}$	-	-0.871^{*}	-	-
	$area_{[NI]}$	-	-0.923^{*}	-0.779^{*}	-
	dep	–	-0.005^{**}	-0.005^{***}	0.005^{**}
	hrec	-	-	-	-
	$seas_{[SP]}$	0.943^{*}	-	-	-
	$seas_{[SU]}$	0.973^{**}	-	-	-
$log(\sigma)$	Intercept		−0.6880	−0.359	−0.343
	$habitat_{[MB]}$		-	-	-
	$area_{[NI]}$		-32.804^{***}	-35.070^{***}	-
	dep		-	-	-
	hrec		-	-	0.059^{*}
	$seas_{[SP]}$		-	1.067^{***}	-
	$seas_{[SU]}$		-	0.265	-

ratio = 0.555, p-value = 0.961), results from the estimated Poisson model suggest a positive effect of the season covariate, with greater abundances in terms of the number of species during spring and summer compared to autumn (Table 1). For the *MaxN* response, an excess of data variability is detected thus the NB distribution is assumed to adjust for overdispersion (dispersion ratio = 3.254, p-value $<$ 0.001). Here the GAMLSS framework is used to model the relation between the mean and the dispersion components of the response distribution and environmental drivers via the appropriate log-link functions. Table 1 shows significant negative effects of habitat, area and depth covariates on the mean of the maximum number of individuals for relevant species, leading to higher abundances in heterogeneous and complex habitats of the Adriatic area and at shallower depths. The geographical factor significantly influences the variability of abundances, particularly regarding the maximum number of individuals with lower dispersion in the northern Ionian Sea. The *Nind* response is also modeled within the GAMLSS framework using the NB assumption to address detected data overdispersion as well (dispersion ratio=3.460, p-value$<$ 0.001). The fitted model reveals significant negative effects for the covariates area and depth. Additionally, a significant negative effect on the variability of count data is associated to the geographical factor playing a significant role in the abundance patterns of selected species together with higher variability of abundances in springer and summer seasons. The GAMLSS framework is also used to esti-

mate the non-negative *1st arrival time* response. A significant effect on mean of species abundance is estimated for the depth covariate suggesting higher abundances associated with lower depths while a positive effect of the recorded hours is associated with higher variability.

5 Conclusion

This study provides insights into species distributions and their connection with the coral systems. Findings underscore how the choice of abundance measure impacts environmental drivers effects on species distribution, with consistent results across different responses. This modelling approach supports further investigation into the potential of baited-lander surveys for studying small-scale distributions of deep-sea marine species and aiding habitat conservation efforts.

References

1. Badger, J.J., Large, S.I., Thorson, J.T.: Spatio-temporal species distribution models reveal dynamic indicators for ecosystem-based fisheries management. ICES J. Mar. Sci. **80**(7), 1949–1962 (2023). https://doi.org/10.1093/icesjms/fsad123
2. Cardinale, B., Primack, R., Murdoch, J.: Conservation Biology. Oxford University Press, Oxford (2019)
3. Chamberlain, S.A., Bronstein, J.L., Rudgers, J.A.: How context dependent are species interactions? Ecol. Lett. **17**, 881–890 (2014). https://doi.org/10.1111/ele.12279
4. Chimienti, G., Bo, M., Taviani, M., Mastrototaro, F.: Occurrence and biogeography of Mediterranean cold-water corals. In: Orejas, C., Jiménez, C. (eds.) Mediterranean Cold-Water Corals: Past, Present and Future, 9, pp. 295–333. Springer, Heidelberg, Germany (2019)
5. D'Onghia, G., et al.: New records of cold-water coral sites and fish fauna characterization of a potential network existing in the Mediterranean Sea. Mar. Ecol. **37**, 1398–1422 (2016). https://doi.org/10.1111/maec.12356
6. Dean, C.B., Lundy, E.R.: Overdispersion. In: Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., Teugels, J.L. (eds.) Wiley StatsRef: Statistics Reference Online (2016). https://doi.org/10.1002/9781118445112.stat06788.pub2
7. R Core Team: RA language and environment for statistical computing, R foundation for statistical. Computing, Vienna, Austria (2023). https://www.R-project.org/
8. Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. Roy. Stat. Soc. Ser. C Appl. Stat. **54**(3), 507–554 (2005). https://doi.org/10.1111/j.1467-9876.2005.00510.x
9. Savini, A., Vertino, A., Marchese, F., Beuck, L., Freiwald, A.: Mapping cold-water coral habitats at different scales within the northern Ionian sea (central Mediterranean): an assessment of coral coverage and associated vulnerability. PLoS ONE **9**(1), e87108 (2014). https://doi.org/10.1371/journal.pone.0087108
10. Ver Hoef, J.M., Boveng, P.L.: Quasi-Poisson vs. negative binomial regression: how should we model over dispersed count data? Ecology **88**(11), 2766–2772 (2007). https://doi.org/10.1890/07-0043.1

Local and Global Economic Policy Uncertainty Influence on US Stock Market Volatility

Vincenzo Candila[1(✉)], Oguzhan Cepni[2], Giampiero M. Gallo[3], and Rangan Gupta[4]

[1] Department of Economics and Statistics, University of Salerno, Fisciano, Italy
vcandila@unisa.it
[2] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark
oce.eco@cbs.dk
[3] Italian Court of Audits (Corte dei conti) and NYU in Florence, Florence, Italy
giampiero.gallo@nyu.edu
[4] Department of Economics, University of Pretoria, Pretoria, South Africa
rangan.gupta@up.ac.za

Abstract. Recently, there has been a growing interest in estimating the global and local Economic Policy Uncertainty (EPU). The resulting global EPU indexes have been used as additional exogenous covariates within mixed-frequency volatility models like the GARCH-MIDAS, usually improving the volatility forecasts. On the other side, the literature lacks analyses where the local (in the sense of state-specific) stock market volatilities are also influenced by local EPU indexes. This paper aims to fill this gap. In particular, we investigate the stock market's daily volatility in ten states of the United States by including a variety of local and global EPU indexes.

Keywords: Financial markets · GARCH-MIDAS · EPU

1 Introduction

Market activity is deeply affected by uncertainty: in particular, the range of return variability that affects asset prices is time-varying and tends to be persistent through time. As a result, asset allocation or risk management rely on forecasting future behavior given the current information set. This so-called phenomenon of volatility clustering [1] is at the basis of a massive literature in financial econometrics, comprising both measurement and modeling of volatility. With relatively recent developments in exploiting the ultra-high frequency data on market prices, measurement is accomplished by building one of the many

The views expressed in the article are those of the authors and do not involve the responsibility of the Corte dei conti.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 321–327, 2025.
https://doi.org/10.1007/978-3-031-64431-3_54

versions of realized volatility, an end-of-day observed variable the dynamics of which can be modeled with either additive [2] or multiplicative errors [3].

In the absence of data on realized volatility, the Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) literature still receives substantial empirical attention, as it simply uses the time series of close-to-close daily returns. The phases of measurement and modeling are accomplished at once: the signal in the conditional mean of the returns is treated as negligible, and the object of interest is the dynamics of the conditional variance, possibly accommodating an asymmetric reaction of volatility to negative returns (as in the Glosten, Jaganathan, and Runkle (GJR, [4]) model). While an autoregressive dependence on past squared returns is common to various GARCH approaches, there has been an increasing awareness about the presence of a slow-moving (i.e., long-run) component around which the short-run dynamics evolve. The GARCH-MIDAS, where MIDAS stands for MIxed Data Sampling, model by [5] provides a very useful setup to accommodate the short- and long-run components. A generalization to the original GARCH-MIDAS comes from [6]. The GARCH-MIDAS model has been recently used in several contributions (see, for instance, [7,8], among others).

In this paper, we aim at forecasting the stock market's daily volatility of ten states of the United States (US) by including a variety of local (i.e., state-specific) and global Economic Policy Uncertainty (EPU) indexes [9]. The ten US states are selected according to a demographic criterion (the top ten by population). The state-level stock market indices derive from the capitalization-weighted index of equities domiciled in a given state. The evaluation of the models under investigation is carried out through the Set of Superior Models (SSM) defined by the Model Confidence Set [10].

2 Empirical Analysis

The framework of this work relies on the general heteroskedastic formulation for the daily log-returns $r_{i,t}$, that is:

$$r_{i,t}|\mathcal{F}_{i-1,t} = \sqrt{\tau_t g_{i,t}}\eta_{i,t}, \tag{1}$$

where i represents the day of the low-frequency period t, $\mathcal{F}_{i-1,t}$ is the information set up to day $i-1$, τ_t is the long-run component varying each period t, $g_{i,t}$ is the short-run component, and $\eta_{i,t}$ is the error term, with $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. In Eq. (1), the short-run component follows a unit-mean GARCH(1,1) process, that is:

$$g_{i,t} = (1-\alpha-\beta) + \alpha r^2_{i-1,t}/\tau_t + \beta g_{i-1,t}, \tag{2}$$

while the long-run is specified as

$$\tau_t = \exp\left\{ m + \theta \sum_{j=1}^{K} \delta(\omega) MV_{t-j} \right\}, \tag{3}$$

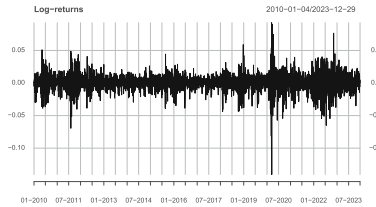
where MV_t is the additional exogenous variable. As said above, the MV_t variable is a variety of EPU indexes.

The state-level dataset for the ten US states has been collected from Bloomberg, covering the time span from January 2010 to December 2023. The log-returns of the ten US states under investigation are plotted in Fig. 1. The monthly EPU indexes used in this work derive from the site https://www.policyuncertainty.com, which provides the *National*, *Local*, and *Composite* EPU indexes for each US state. These three EPU indexes are described in the first three rows of Table 1 while Fig. 2 illustrates the graphical patterns for the ten US states. In addition to the state-specific EPU indexes, [9] also provide a global US EPU index. All these EPU indexes are based on textual analysis. Besides the four EPU indexes, we also include seven other EPU indexes in the GARCH-MIDAS model as low-frequency variables, as described in Table 1. Overall, we estimate eleven GARCH-MIDAS models, using a rolling window of eleven years and a re-estimation frequency of thirty days. The out-of-sample period is from January 2021 to December 2023 (750 daily observations).

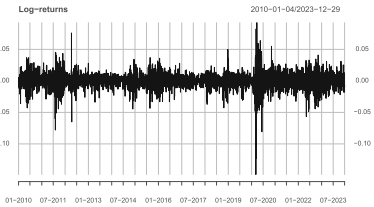
Table 2 reports the Mean Squared Errors (MSE) of the eleven models involved in our analysis. The shades of gray indicate inclusion in the SSM at 25% significance level. Remarkably, the state-specific EPU helps in forecasting the daily volatility. If the *National* EPU and the average-by-state of the three state-specific EPU indexes enter the SSM for eight states out of ten, the global US EPU index enters the SSM only for six states.

Table 1. Low-frequency variables

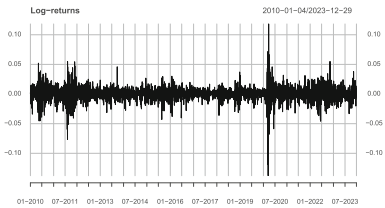
Label	Description
Nat.	Uncertainty from national policy-related sources
Loc.	Uncertainty from state and local policy issues
Comp.	Uncertainty from national and state-specific sources
Avg.	Uncertainty obtained averaging Nat., Loc. and Comp.
Nat. Avg.	Uncertainty obtained averaging all the Nat. observations
Loc. Avg.	Uncertainty obtained averaging all the Loc. observations
Comp. Avg.	Uncertainty obtained averaging all the Comp. observations
Neigh. Nat.	Uncertainty for a State i obtained averaging all the Nat. observations of the bordering states
Neigh. Loc.	Uncertainty for a State i obtained averaging all the Loc. observations of the bordering states
Neigh. Comp.	Uncertainty for a State i obtained averaging all the Comp. observations of the bordering states
Glob.	Global US uncertainty



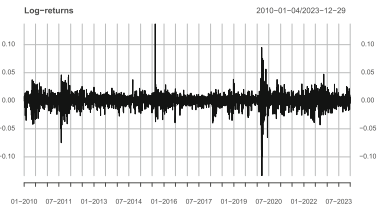
(a) California



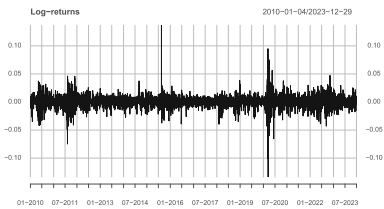
(b) Texas



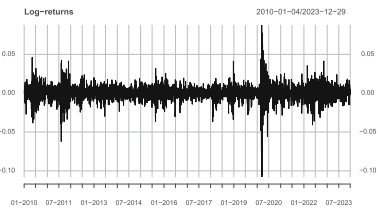
(c) Florida



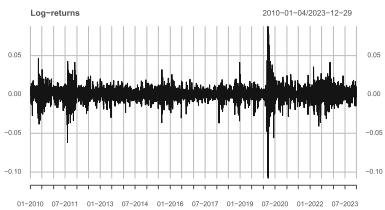
(d) New York



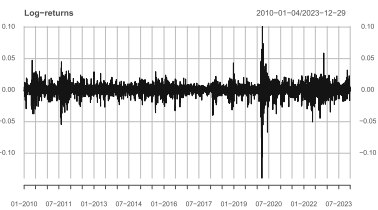
(e) Pennsylvania



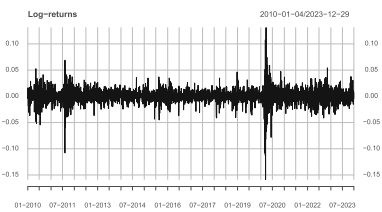
(f) Illinois



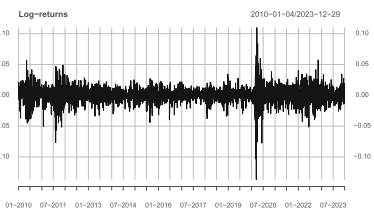
(g) Ohio



(h) Georgia

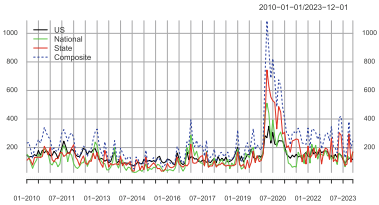


(i) North Carolina

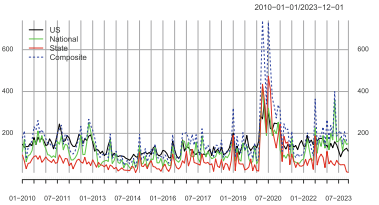


(j) Michigan

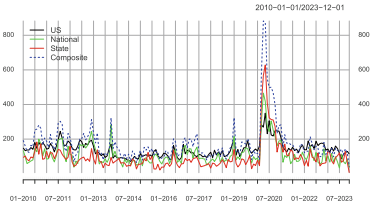
Fig. 1. Daily log-returns



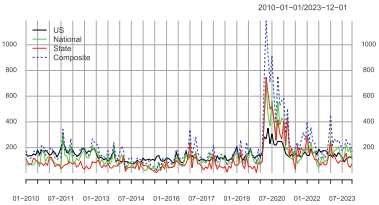
(a) California



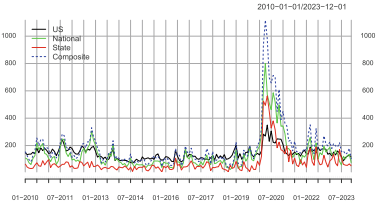
(b) Texas



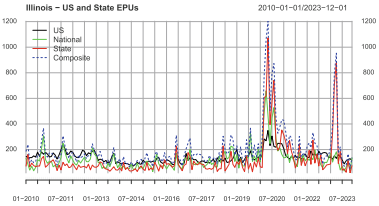
(c) Florida



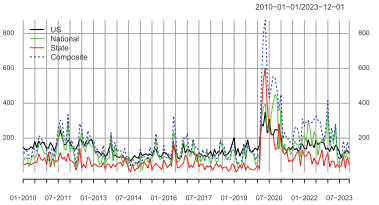
(d) New York



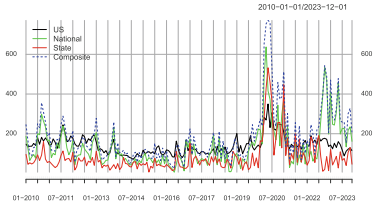
(e) Pennsylvania



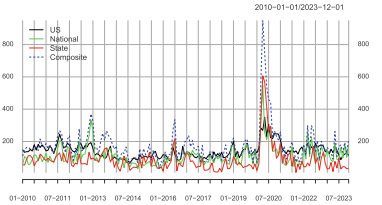
(f) Illinois



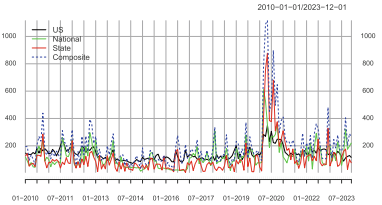
(g) Ohio



(h) Georgia



(i) North Carolina



(j) Michigan

Fig. 2. Global and State-specific EPUs

Table 2. Out-of-sample volatility evaluation

State/EPU	GARCH-MIDAS										
	Nat.	Loc.	Comp.	Avg.	Nat. Avg.	Loc. Avg.	Comp. Avg.	Neigh. Nat.	Neigh. Loc.	Neigh. Comp.	Glob.
California	1.333	1.352	1.351	1.339	1.341	1.357	1.356	1.337	1.384	1.393	1.348
Texas	0.487	0.492	0.489	0.487	0.490	0.502	0.495	0.488	0.494	0.494	0.494
Florida	0.379	0.385	0.385	0.384	0.38	0.381	0.379	0.379	0.377	0.378	0.379
New York	0.241	0.239	0.246	0.240	0.242	0.282	0.241	0.24	0.244	0.241	0.244
Pennsylvania	0.289	0.289	0.288	0.288	0.288	0.289	0.288	0.288	0.287	0.287	0.296
Illinois	0.159	0.157	0.156	0.155	0.159	0.159	0.158	0.159	0.157	0.158	0.160
Ohio	0.177	0.183	0.178	0.178	0.179	0.179	0.179	0.178	0.183	0.179	0.179
Georgia	0.354	0.358	0.354	0.354	0.356	0.359	0.358	0.358	0.359	0.359	0.355
North Carolina	0.466	0.466	0.463	0.464	0.462	0.465	0.472	0.465	0.485	0.465	0.463
Michigan	0.587	0.642	0.583	0.792	0.584	0.585	0.584	0.591	0.582	0.584	0.583

Notes: The table reports the MSE of the GARCH-MIDAS. The low-frequency variables of the GARCH-MIDAS models are illustrated in Table 1. Sample period: January 2021 to December 2023 (750 observations). Shades of gray indicate inclusion in the SSM at 25% significance level.

References

1. Engle, R.F.: Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica **50**(4), 987–1008 (1982). https://doi.org/10.2307/1912773
2. Corsi, F.: A simple approximate long-memory model of realized volatility. J. Financ. Econom. **7**(2), 174–196 (2009). https://doi.org/10.1093/jjfinec/nbn005
3. Cipollini, F., Gallo, G.M., Otranto, E.: Realized volatility forecasting: robustness to measurement errors. Int. J. Forecast. **37**(1), 44–57 (2021). https://doi.org/10.1016/j.ijforecast.2020.02.009
4. Glosten, L.R., Jaganathan, R., Runkle, D.E.: On the relation between the expected value and the volatility of the nominal excess return on stocks. J. Finance **48**(5), 1779–1801 (1993). https://doi.org/10.1111/j.1540-6261.1993.tb05128.x
5. Engle, R.F., Ghysels, E., Sohn, B.: Stock market volatility and macroeconomic fundamentals. Rev. Econ. Stat. **95**(3), 776–797 (2013). https://doi.org/10.1162/REST_a_00300
6. Amendola, A., Candila, V., Gallo, G.M.: On the asymmetric impact of macro-variables on volatility. Econ. Model. **76**, 135–152 (2019). https://doi.org/10.1016/j.econmod.2018.07.025
7. Salisu, A.A., Gupta, R., Demirer, R.: Global financial cycle and the predictability of oil market volatility: evidence from a GARCH-MIDAS model. Energ. Econ. **108**, 105934 (2022). https://doi.org/10.1016/j.eneco.2022.105934

8. Segnon, M., Gupta, R., Wilfling, B.: Forecasting stock market volatility with regime-switching GARCH-MIDAS: the role of geopolitical risks. Int. J. Forecast. **40**(1), 29–43 (2024). https://doi.org/10.1016/j.ijforecast.2023.04.007
9. Baker, S.R., Bloom, N., Davis, S.J.: Measuring economic policy uncertainty. Q. J. Econ. **131**(4), 1593–1636 (2016). https://doi.org/10.1093/qje/qjw024
10. Hansen, P.R., Lunde, A., Nason, J.M.: The model confidence set. Econometrica **79**(2), 453–497 (2011). https://doi.org/10.3982/ECTA5771



An EWMA Control Chart for Multivariate Functional Data

Christian Capezza[1(✉)], Giovanna Capizzi[2], Fabio Centofanti[1], Antonio Lepore[1], and Biagio Palumbo[1]

[1] Department of Industrial Engineering, University of Naples Federico II, Napoli, Italy
christian.capezza@unina.it

[2] Department of Statistical Sciences, University of Padua, Padova, Italy

Abstract. This paper introduces a novel functional control chart that extends the classical exponentially weighted moving average (EWMA) control chart from univariate scalar data to multivariate functional data. Through extensive Monte Carlo simulations, the proposed control chart demonstrates superior performance compared to competing monitoring schemes. Its practical utility is illustrated through a case study in automotive manufacturing, where it is employed to monitor the quality of resistance spot welding processes by analyzing dynamic resistance curves.

Keywords: Functional Data Analysis · Statistical Process Control

1 Introduction

Profile monitoring [8] aims to build control charts for the statistical process monitoring (SPM) of quality characteristics in the form of one or multiple functional variables [2,4,6]. Most of the existing works on this topic focus on Shewhart-type control charts, which allow the process state to be assessed only on the information acquired at the current point in time, without considering previous data points, which may contain decisive information about process condition trends.

The exponentially weighted moving average (EWMA) control chart is instead known to overcome this limitation by evaluating the current state of the process through a statistic calculated as a weighted average of the observation at the current point in time and the statistic itself at the previous point. However, only a few works focus on EWMA control charts for profile monitoring. In particular, [7] developed a functional EWMA control chart for both univariate and multivariate functional data, while [9] move forward with the integration of the EWMA scheme with the multichannel functional principal component analysis.

The main limitation of these works is related to the choice of the weighting parameter of the control chart, usually denoted as λ. Typically, a small value of λ is recommended to detect small mean shifts, while a large value is preferred instead to identify large shifts. However, in most real cases, there is no prior

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 328–333, 2025.
https://doi.org/10.1007/978-3-031-64431-3_55

information available on the magnitude of the shift that may actually occur in the OC scenario.

Therefore, in this paper, we illustrate a new method for the monitoring of a process through observations of a multivariate functional quality characteristic, which is named adaptive multivariate functional EWMA (AMFEWMA) control chart and extends the underlying idea of Capizzi and Masarotto [5] to the multivariate profile monitoring setting.

2 Methodology

The methodology is based on [1]. Let us consider a sequence of independent realizations in $\mathbb{H}$ of the random vector $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{np})^T$, $n = 1, 2, \ldots$ with zero mean. The AMFEWMA statistic $\boldsymbol{Y}_n = (Y_{n1}, \ldots, Y_{np})^T$, is defined as

$$\boldsymbol{Y}_n(t) = \boldsymbol{Y}_{n-1}(t) + \boldsymbol{\Lambda}_n(t)\boldsymbol{E}_n(t), \quad t \in \mathcal{T}, n = 1, 2, \ldots, \tag{1}$$

where $\boldsymbol{E}_n(t) = (E_{n1}(t), \ldots, E_{np}(t))^T = \boldsymbol{X}_n(t) - \boldsymbol{Y}_{n-1}(t)$, while $\boldsymbol{\Lambda}_n(t)$ is a diagonal matrix with j-th diagonal element $w(E_{nj}(t)) = \eta(E_{nj}(t))/E_{nj}(t)$ and $\eta(E_{nj}(t))$ is a score function evaluated at the error $E_{nj}(t)$, $j = 1, \ldots, p$. The AMFEWMA statistic $\boldsymbol{Y}_n$ is a weighted average of current observation $\boldsymbol{X}_n$ and the value of the charting statistic at the previous time point $\boldsymbol{Y}_{n-1}$, as in a conventional EWMA chart, but with weights $w(E_{nj}(t))$ changing over time. The score function η extends the score function for a univariate scalar quality characteristic, proposed by [5], to the multivariate functional data setting

$$\eta(E_{nj}(t)) = \begin{cases} E_{nj}(t) + (1-\lambda)C_j(t) & \text{if } E_{nj}(t) < -C_j(t) \\ \lambda E_{nj}(t) & \text{if } |E_{nj}(t)| \leq C_j(t) \\ E_{nj}(t) - (1-\lambda)C_j(t) & \text{if } E_{nj}(t) > C_j(t) \end{cases}, \quad t \in \mathcal{T}, \tag{2}$$

where $\lambda \in (0, 1]$ and $C_j(t) = k\sigma_j(t)$, with k denoting a real positive constant and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^T$ representing the standard deviation function of $\boldsymbol{X}_n$. When $\boldsymbol{X}_n$ approaches $\boldsymbol{Y}_{n-1}$, i.e., $E_{nj}(t)$ tends to zero, the AMFEWMA chart is expected to perform like the MFEWMA control chart, while it is expected to behave like a Shewhart control chart when $|E_{nj}(t)|$ is large.

The monitoring scheme is based on the functional extension of the classical Hotelling's statistic, calculated as the following function of the AMFEWMA statistic

$$V_n^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\int_{\mathcal{T}}\int_{\mathcal{T}} Y_{ni}(s)K_{ij}^*(s,t)Y_{nj}(t)dsdt, \tag{3}$$

where the function K_{ij}^* is defined as

$$K_{ij}^*(s,t) = \sum_{l=1}^{L}\frac{1}{\rho_l}\psi_{li}(s)\psi_{lj}(t), \quad s, t \in \mathcal{T}, i, j = 1, \ldots, p, \tag{4}$$

and $\boldsymbol{\psi}_l = (\psi_{l1}, \ldots, \psi_{lp})^T$ denotes the eigenfunctions corresponding to the eigenvalues ρ_l of the asymptotic covariance $\boldsymbol{C}(s,t)$ of $\boldsymbol{Y}_n$. The number of principal components L is usually chosen such that the retained principal components explain a given fraction of the total variability.

The AMFEWMA control chart is designed to signal when the statistic V_n^2 given in Equation (3) exceeds the upper control limit h chosen as the smallest value achieving an in-control average run length, denoted as ARL_0, and calculated through a bootstrap procedure. We remind to [1] for additional details on the calculation of the control limits and the selection of the control chart parameters λ and k introduced above.

In the Phase II monitoring, given the current sequence $\boldsymbol{X}_n$, the AMFEWMA statistic $\boldsymbol{Y}_n$ and the corresponding monitoring statistic V_n^2 are obtained from Eqs. (1) and (3), respectively, according to the asymptotic covariance function $\boldsymbol{C}(s,t)$ estimated in Phase I. An alarm signal is issued if V_n^2 exceeds the control limit h calculated in Phase I.

3 Simulation Study

The effectiveness of the proposed AMFEWMA control chart in detecting shifts in the means of multivariate functional data is assessed through an extensive Monte Carlo simulation analysis. The data generation procedure follows the approach outlined in [4] and provides the typical functional patterns observed in dynamic resistance curves within the automotive manufacturing process, specifically in resistance spot welding. Multivariate functional quality data consisting of 5 components ($p = 5$) are generated. Out-of-control (OC) observations simulate a splash weld phenomenon, also referred to as expulsion, caused by excessive welding current passing through the electrodes [10]. Six severity levels ($SL \in \{0, 1, 2, 3, 4, 5, 6\}$) are investigated, where $SL = 0$ indicates that the process is in control. These severity levels gauge the degree to which profiles in the Phase II dataset deviate from their Phase I counterparts.

The proposed control chart is compared with two alternative methods: the MFEWMA control chart for multivariate functional data, introduced in [9], with the Phase II monitoring scheme having a weighting parameter $\lambda \in \{0.1, 0.5\}$; and the traditional multivariate Shewhart control chart, based on the Hotelling T^2 statistic on coefficients derived from MFPCA [3], referred to as SHEWHART.

In each scenario and for each severity level (SL), 30 simulation runs are conducted. Each run involves a Phase I sample comprising 2500 observations. The Phase II dataset consists of 200 sequences of independently and identically distributed (i.i.d.) observations, with shift locations set to 100.

The performance evaluation of the AMFEWMA method and its competitors is carried out by assessing the estimated average run length ($\widehat{ARL}$) across different severity levels (SL). Ideally, under in-control (IC) conditions, $\widehat{ARL}$ should closely approximate the fixed value of ARL_0, set at 20, while it should be minimized when the process is out of control (OC).

In Fig. 1, the estimated $\widehat{ARL}$ values are presented for the SHEWHART, MFEWMA (with $\lambda \in \{0.1, 0.5\}$), and the proposed control chart, denoted as AMFEWMA*, within the simulation study.

Among the competing methods, the SHEWHART control chart demonstrates superior performance for larger SLs, while the MFEWMA with smaller λ values emerges as the preferred approach for smaller SLs. The $\widehat{ARL}$ values achieved by the proposed control chart indicate satisfactory performance across all considered OC conditions.

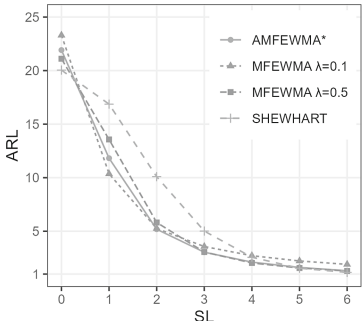


Fig. 1. The estimated ARL values for the SHEWHART, MFEWMA (with $\lambda \in \{0.1, 0.5\}$) and the AMFEWMA* control charts.

4 Case Study

To illustrate the practical utility of the proposed AMFEWMA control chart, we present a case study focusing on the statistical monitoring of a resistance spot-welding process in automotive body-in-white manufacturing. The data under analysis are provided by Centro Ricerche Fiat and were collected during lab tests conducted at the Mirafiori Factory on various bodies of the same car model. Each body exhibits numerous spot welds and varying characteristics, such as metal sheet thickness, material, and welding time. The case study specifically examines a set of 10 spot welds performed with the same welding gun on different car bodies. Thus, for each car body, we evaluate the multivariate functional quality characteristic represented by the vector of 10 Dynamic Resistance Curves (DRCs) corresponding to the same 10 spot welding points.

The Phase I sample consists of profiles observed immediately after electrode renewal, a routine procedure aimed at neutralizing wear effects known to impact the quality of the final welded point, as mentioned in the introduction. The remaining observations are divided into three datasets, each corresponding to incremental levels of electrode wear: wear levels 1, 2, and 3, respectively. These datasets serve as Phase II observations to assess the performance of the proposed chart concerning different levels of shift associated with electrode wear.

Figure 2 depicts the application of the AMFEWMA* approach to the Phase II monitoring of the three datasets corresponding to increasing wear levels. In each plot, the first 100 observations are randomly sampled from the Phase I sample, while the Phase II observations for each wear level are shown to the right of the dashed vertical line. The figure demonstrates that the AMFEWMA* scheme effectively detects mean shifts across all severity levels.

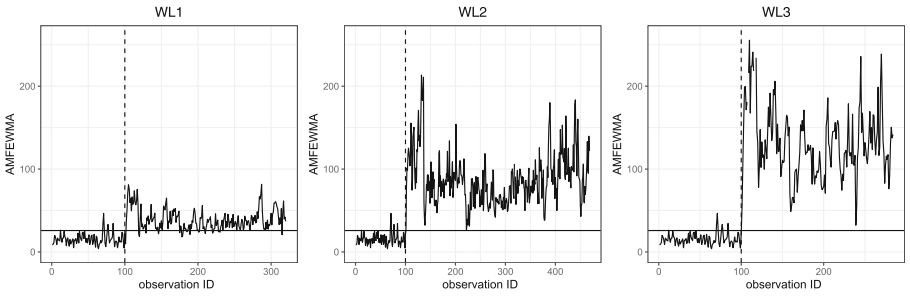


Fig. 2. AMFEWMA* control chart under wear levels (WL) 1, 2, and 3. In each plot, the Phase II observations are reported after the dashed vertical line.

5 Conclusions

A novel control chart designed for the statistical process control of multivariate functional data is introduced, termed the adaptive multivariate functional exponentially weighted moving average control chart (AMFEWMA). Through an extensive Monte Carlo simulation, the performance of the AMFEWMA chart is compared with that of two competing methods. The results demonstrate the superiority of AMFEWMA in correctly identifying out-of-control (OC) scenarios when the magnitude of the mean shift is unknown. Specifically, for small shifts, the AMFEWMA behaves similarly to the conventional MFEWMA control chart, while for larger shifts, it aligns with the performance of the Shewhart control chart. The practical usefulness of the proposed method has been demonstrated through a case study in the automotive industry.

Acknowledgements. This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GeneraFonEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, The activity of B. Palumbo is supported by the MOST - Sustainable Mobility National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 - D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

1. Capezza, C., Capizzi, G., Centofanti, F., Lepore, A., Palumbo, B.: An adaptive multivariate functional EWMA control chart. Accepted for publication on J. Qual. Technol. (2024). https://doi.org/10.1080/00224065.2024.2383674
2. Capezza, C., Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional regression control chart for monitoring ship CO2 emissions. Qual. Reliab. Eng. Int. **38**(3), 1519–1537 (2022)
3. Capezza, C., Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: funcharts: control charts for multivariate functional data in R. J. Qual. Technol. **55**(5), 566–583 (2023)
4. Capezza, C., Centofanti, F., Lepore, A., Palumbo, B.: Robust multivariate functional control charts. Technometrics **66**(4), 531–547 (2024). https://doi.org/10.1080/00401706.2024.2327346
5. Capizzi, G., Masarotto, G.: An adaptive exponentially weighted moving average control chart. Technometrics **45**(3), 199–207 (2003)
6. Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional regression control chart. Technometrics **63**(3), 281–294 (2021)
7. Fassò, A., Toccu, M., Magno, M.: Functional control charts and health monitoring of steam sterilizers. Qual. Reliab. Eng. Int. **32**(6), 2081–2091 (2016)
8. Noorossana, R., Saghaei, A., Amiri, A.: Statistical Analysis Profile Monitoring. Wiley (2011)
9. Ren, H., Chen, N., Wang, Z.: Phase-II monitoring in multichannel profile observations. J. Qual. Technol. **51**(4), 338–352 (2019)
10. Xia, Y.J., Su, Z.W., Li, Y.B., Zhou, L., Shen, Y.: Online quantitative evaluation of expulsion in resistance spot welding. J. Manuf. Process. **46**, 34–43 (2019)

Low-Income Earners: New Possibilities of Study from Administrative Archives

Monica Carbonara[1(✉)] and Sara Casacci[2]

[1] Italian National Institute of Statistics, Bari, Italy
mocarbon@istat.it
[2] Italian National Institute of Statistics, Roma, Italy
casacci@istat.it

Abstract. The phenomenon of the "working poor" has gained attention both in the political debate and in socio-economic research over the last years, due to its general increase. This paper proposes an analysis of the characteristics of low-income employees by exploiting the information potential of integrated administrative archives. Since the data contain information on income and employment of all household members, it is possible to analyse economic disadvantage at both individual and household level. In particular, the paper describes the relationship that exists between the probability of being a low-income employee in Italy and a set of socio-demographic variables. An analysis of the phenomenon at a low territorial level is also presented, with the aim of providing useful elements to identify particularly vulnerable areas.

Keywords: low-income earners · administrative archives · logistic regression · territorial analysis

1 The Theoretical Framework

In recent years, the phenomenon of "poor workers" has grown, i.e. those who, despite working, earn an income below the poverty threshold. The problem of the limited quantity of work, together with its inadequate quality (low wages, few contribution guarantees, irregular careers, etc.) seem to identify a new alarming problem of social policy which entails a significant risk of individual poverty for workers and increases that of the entire family unit. In the public debate, working poverty is often linked to insufficient wages while it is the result of a process that goes well beyond wages and concerns working times (i.e. how many hours one usually works per week and how many weeks one works in the over the course of a year), family composition (and in particular how many people receive an income within the household) and the redistributive role of the State. At an individual level, in fact, the risk of low wages is particularly high for workers employed only a few months a year, for part-time workers and for self-employed workers. At family level, the composition of the household and the number of income earners are also added to these risk factors. There are numerous studies on the topic of income distribution and poverty, which, however, have mostly used the results of sample surveys, with the consequent limits of significance of the data when the level of detail, thematic or territorial, it is very fine.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 334–340, 2025.
https://doi.org/10.1007/978-3-031-64431-3_56

This work aims to analyse low-income employment in Italian municipalities, to identify the main determinants of the phenomenon and the areas particularly exposed, by exploiting the information potential of integrated administrative archives.

In line with the definitions used at European level, the poverty threshold is set at 60% of the median income[1]. Nevertheless, it is not appropriate to use the term 'poverty' in relation to administrative data. The following definitions are adopted:

- employee: individual, who in the microdata collection has an administrative sign of employment in October 2019, receives income from employment, and has signs of work activity for at least six months of the year.
- Low-income employee: employee, as defined in the previous point, with income from employment less than or equal to 60% of the median income from employment.
- Household: group of persons linked by ties of marriage, kinship, affinity, adoption, guardianship or cohabitation and registered in the municipal registers.
- Low-income household (economic disadvantaged household): household with gross equivalent family income lower than 60% of the median gross equivalent income.

2 Data and Method

2.1 Data

The application described in this paper deals with a collection of microdata produced as part of the ArchIMEDe Project of the Italian National Institute of Statistics to study the socio-economic conditions of individuals and households in Italy at the municipal level [4]. Data refers to 2019. This microdata collection is based on the integration of several administrative sources and statistical registers. It collects information on socio-demographic characteristics, occupation, education and income (employee income and total income). In particular, the employment variables derive from the integration of multiple Social Security sources, income variables from the integration of the Tax Return Register with other sources such as the Central Register of Pensioners and the Social Security Benefits Register. The information derived from administrative sources allows a complete coverage of the reference populations (the universe of individuals and households resident in Italy). It should be noted that there are some informational gaps and comparability with surveys is not guaranteed. In particular, microdata only cover regular employment and provide income before tax (therefore, income is not affected by the redistributive effects of taxation). Moreover, income may suffer from under coverage, e.g. due to tax evasion or low compliance. Since information on employees is considered more reliable, the analysis is carried out on this category of workers.

2.2 Method

In order to describe the relationship that exists between the probability of being an employee with an income less than or equal to 60% of the median and a set of socio-demographic variables, a logistic regression model was used. For the selection of the

[1] Eurostat sets the risk-of-poverty threshold at 60% of the national median equivalised disposable income (after social transfers).

regressors, a stepwise procedure was performed. It builds the model iteratively, with variables gaining or losing significance being added or removed at each step. The parameters were estimated using the maximum likelihood method, according to which the preferred values of the parameters are those that maximize the probability of obtaining the observed data (log-likelihood function). To analyse the phenomenon with low territorial detail, we used thematic maps with raw values grouped into quartiles ("quartile maps"). This solution is commonly adopted to visualize the distribution territorial of a phenomenon and to identify areas of greater or lesser intensity [3].

3 Results

3.1 Who Are the Italian Low-Income Employees?

According to our data, around 14 million people have an administrative sign of employee work in the month of October and have work signals for at least six months in 2019. The 79.1% of them are outside the area of economic disadvantage (Fig. 1), while 6.4% show an economic disadvantage at both individual and household level.

		Employee	
		with income from employment greater than 60% of the median	with income from employment less than or equal to 60% of the median
Household	with total income greater than 60% of the median	79.1	11.3 (A)
	with total income less than or equal to 60% of the median	3.2 (B)	6.4 (C)

Fig. 1. Relationship between employees' income and households' income (percentage)

The results of the logistic regression model show that type of employment contract, sex, level of education and Region of residence are the most important independent variables. Foreign citizenship, university student status, number of household members and marital status are also variables that significantly improve the fit of the model. The three tests (Likelihood Ratio, Score and Wald) to test the null hypothesis that all explanatory variables have a regression coefficient equal to zero are significant, so that the null hypothesis can be rejected and the conclusion can be drawn that at least one of the model's regression coefficients is different from zero. Furthermore, the hypothesis tests conducted for the variables included in the model indicate that all variables contribute to a significant improvement in fit. Table 1 reports the parameter estimates and the corresponding standard errors, the test statistics and the p-value, used to test the null hypothesis that the regression coefficient of each individual predictor is equal to zero, given the other explanatory variables of the model. The conditions that contribute to increase the probability of having an income from employment less than or equal to 60% of the median are: being resident in a southern region (especially Campania, Sicily, Calabria), being a woman, being a foreigner, having an educational qualification lower

than a diploma, having a precarious work. The results obtained are quite consistent with other works on the topic [1, 2, 5].

Table 1. Analysis of Maximum Likelihood Estimates.

Parameter		Estimate	Standard error	Wald Chi-square	Pr > Q^2
Intercept		0.471	0.0065	5.301	<.0001
Region of residence	Piemonte	−0.448	0.0033	18.687	<.0001
Region of residence	Valle d'Aosta	−0.567	0.018	1.005	<.0001
Region of residence	Lombardia	−0.501	0.0024	43.688	<.0001
Region of residence	Bolzano	−0.875	0.0083	11.178	<.0001
Region of residence	Veneto	−0.424	0.0023	20.051	<.0001
Region of residence	Friuli-Venezia Giulia	−0.462	0.0057	6.535	<.0001
Region of residence	Liguria	−0.117	0.0050	535	<.0001
Region of residence	Emilia-Romagna	−0.605	0.0033	34.607	<.0001
Region of residence	Toscana	−0.200	0.0033	3.698	<.0001
Region of residence	Umbria	−0.088	0.0063	192	<.0001
Region of residence	Marche	−0.142	0.0049	858	<.0001
Region of residence	Lazio	0.191	0.0027	4.885	< .0001
Region of residence	Abruzzo	0.218	0.0052	1.729	<.0001
Region of residence	Molise	0.601	0.0105	3.260	<.0001
Region of residence	Campania	0.738	0.0029	67.051	< .0001
Region of residence	Puglia	0.612	0.0033	35.452	<.0001

(*continued*)

Table 1. (*continued*)

Parameter		Estimate	Standard error	Wald Chi-square	Pr > Q^2
Region of residence	Basilicata	0.500	0.0078	4.183	<.0001
Region of residence	Calabria	0.876	0.0045	35.309	<.0001
Region of residence	Sicilia	0.782	0.0031	64.922	<.0001
Region of residence	Sardegna	0.316	0.0045	4.629	<.0001
Sex	Male	−0.640	0.0000	564.537	<.0001
Marital status	Unmarried	0.046	0.0015	979	<.0001
Marital status	Married	−0.130	0.0013	9.730	<.0001
Citizenship	Italian	−0.336	0.0013	65.903	<.0001
Number of household's members		0.084	0.0006	17.250	<.0001
Highest level of education	Up to lower secondary school certificate	0.944	0.0038	61.189	<.0001
Highest level of education	Up to upper secondary school certificate	0.211	0.0038	3.127	< .0001
Highest level of education	Tertiary	−0.267	0.0109	612	<.0001
Age		−0.032	0.0001	141.643	<.0001
University student	No	−0.369	0.0029	16.124	<.0001
Precarious contract	No	−0.797	0.0010	629.508	<.0001

3.2 A Territorial Analysis

From the analysis of municipal data, the evident territorial concentration of the phenomenon in the regions of Southern Italy emerges (Fig. 2). In fact, in most southern municipalities three out of ten workers have an income lower than or equal to 60% of the median or live in household with an income below the threshold. Indeed, 70% of

the municipalities in Southern Italy belong to the fourth quartile of the municipal distribution. The similar percentage for the municipalities of the North is 2%, 10.8% for the municipalities of the Centre.

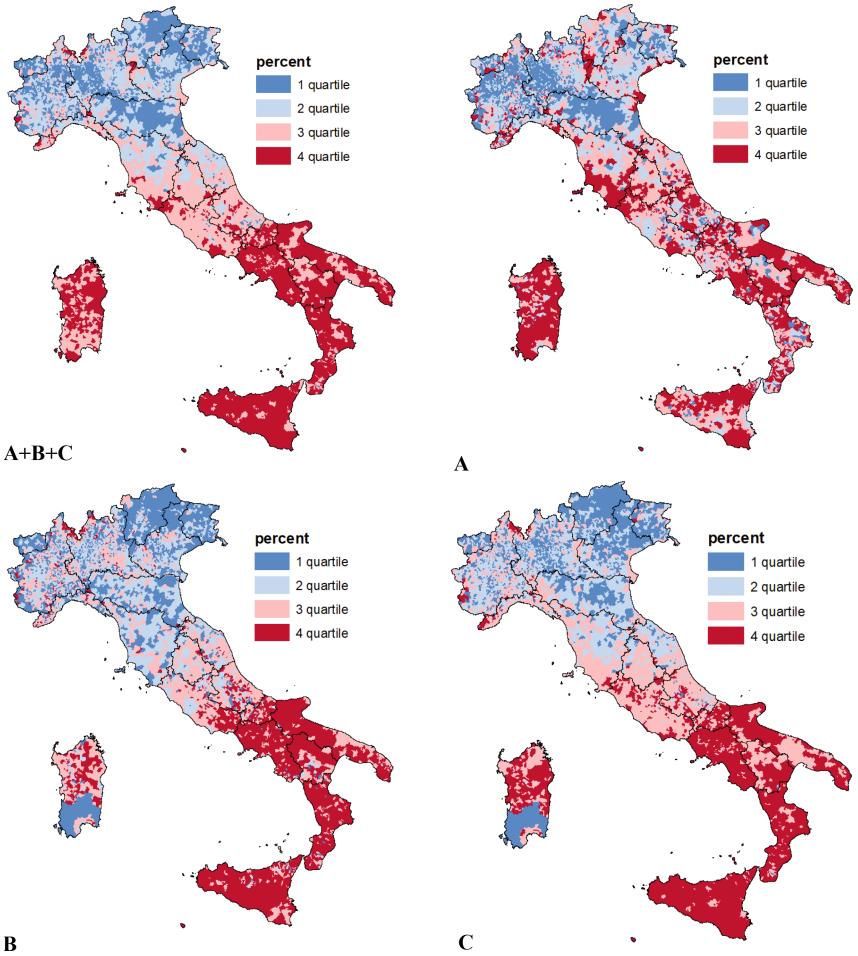


Fig. 2. Incidence of low-income employees in non-low-income households (A), non-low-income employees in low-income households (B), low-income employees in low-income households (C) and low-income employees or employees living in low-income households (A + B + C). Quartile maps.

Restricting the analysis to low income employees living in economic disadvantaged households (Cartogram C of Fig. 2), the ratio drops to four households out of ten, but the greater concentration in the southern regions is confirmed. As regards, instead, low income employees living in households with an income above the threshold (situation of less disadvantage, in which the household plays a role of social protection), the North-South gap diminishes (Cartogram A of Fig. 2). In fact, for this population, 53.3%

of the municipalities in the South belong to the last quartile, compared to 7.9% of the municipalities in the North and 27.8% of the municipalities in the Centre[2]. It is interesting to note the presence of areas, even in the South, in which the role of the family mitigates the scarcity of employment income.

4 Conclusion

The analysis of low-income employment proposed in this paper underlines the very great potential of integrated administrative data, even if they are currently limited. The richness and multidimensionality of this kind of data made it possible to carry out a descriptive study of the phenomenon in Italy. We performed a multivariate analysis, describing the relationship between the probability of being a low-income employee and a set of socio-demographic variables. The data also provide an opportunity to experiment with new operational definitions. In particular, the inclusion of equivalised income allowed us to segment the population and delineate three further conceptual categories: (1) low-income employees living in a household with economic deprivation, (2) non low-income employees living in a household with economic deprivation (3) low-income employees living in a household without economic disadvantage. Finally, as the data allows for granularity at the sub-regional level, a territorial analysis is presented. This aspect is particularly useful for local decision-makers. The results show the strong territorial variability and confirm the North-South dualism. However, the analysis of the new categories shows a more differentiated picture and highlights the presence of areas where the role of the family mitigates the scarcity of employment income and vice versa.

References

1. Bavaro, M.: Is working enough? a study on low-paid workers in Italy. WorkINPS papers (2022)
2. Bovini, G., Ciani, E., De Philippis E., Romano, S.: Labour income inequality and in-work poverty: a comparison between euro area countries. Occasional Papers Banca d'Italia (2023)
3. Cariello, S., Carbonara, M.: Working poor nel Nord ovest. In Verrecchia, F. (ed.).: Dati amministrativi, metodi e statistiche per le politiche territoriali, pp. 85–92. Franco Angeli, Milano (2019)
4. Garofalo, G.: Il Progetto ARCHIMEDE obiettivi e risultati sperimentali. Istat Working Paper (2014)
5. Raitano, M., Jessoula M., Pavolini, E., Natili, M.: ESPN Thematic Report on Inwork poverty – Italy. European Social Policy Network (ESPN), Brussels, European Commission (2019)

[2] A high incidence of the phenomenon is also observed in some municipalities bordering France and Switzerland, due to the distorting effect of cross-border work. In fact, the incomes of cross-border workers are not included in the ArchIMEDe collection of micro-data.

Inequalities Among Older People: Exploring Differences in Deprivation Patterns

Maria Carella, Roberta Misuraca(✉), and Anna Paterno

University of Bari Aldo Moro, Bari, Italy
{maria.carella1,roberta.misuraca,anna.paterno}@uniba.it

Abstract. In the last twenty years, a large body of literature has explored socioeconomic inequalities by focusing on the imprecise nature of deprivation. From a methodological perspective, numerous approaches have been used to measure differences in the concept of deprivation among different target groups. However, despite the significance of this issue, no previous research has evaluated the attributes resulting from the combination of the variety and intensity of different deprivation indicators among the elderly in Italy. This exploratory study uses a Latent Class Analysis of the data from the 2018 Italian Multipurpose Survey on Households to identify three distinct profiles of deprivation based on different degrees of hardship related to specific indicators. This innovative approach enables us to better understand the specific interplay between varying degrees of severity across different deprivation items. Our results show that food deprivation is more uniformly distributed within the older population. In contrast, housing and material deprivation are more distinctly characterised, particularly among those aged over 75, those living in Southern Italy, and households with traditional structures. This calls for further research, highlighting the need for nuanced investigations that consider the diverse forms of deprivation across subpopulations.

Keywords: Material deprivation · Older population · LCA · Italy

1 Introduction and Theoretical Background

Numerous studies over the past twenty years have addressed socioeconomic inequalities by analysing the imprecise nature of deprivation, frequently adopting a comparative perspective (Kim 2016). These studies have shown that inequalities derived from the multiple dimensions of deprivation are a socially constructed phenomenon that can be analysed according to the nature of the sub-populations and the social and/or geographical context (Bonarini and d'Ercole 2006; Alkire and Foster 2011).

From a methodological perspective, numerous approaches have been used to measure differences in deprivation between different target groups (Navarro and Ayala 2008; Panek 2010). However, despite the significance of this issue, to the best of our knowledge no previous research has explored the patterns of deprivation resulting from the combination of its different dimensions among the elderly in Italy.

This study aims to update the state-of-the-art on this subject by investigating inequalities among older people in Italy through a combination of the variety and intensity of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 341–346, 2025.
https://doi.org/10.1007/978-3-031-64431-3_57

different deprivation indicators. Our goal is to further explore the specific interplay between varying degrees of deprivation across different items. Although insufficient to be directly linked to a conventional concept of poverty, these items collectively contribute to a nuanced understanding of deprivation.

We have conducted a Latent Class Analysis (LCA) using data from the Multipurpose Survey on Households, Aspects of Daily Life (Italian National Institute of Statistics - ISTAT, 2018) to identify different profiles of deprivation based on their patterns/level of hardship involving specific dimensions. We provide a descriptive analysis of the dynamics of deprivation within the older population. In a second step, we employ a probit regression analysis to identify various individual characteristics of older people that are more likely to be correlated with the likelihood of experiencing deprivation across different dimensions, while accounting for factors such as income sources and physical limitations. This allows us to clarify the role played by crucial attributes such as gender, age, and education level in the probability of deprivation. In so doing, we seek to answer the following research questions:

RQ1: Which dimension of deprivation is more likely to be experienced by older individuals when differentiating by age group?

RQ2: Which sociodemographic characteristics of older people have a stronger association with the different dimensions of deprivation?

This seminal work seeks to expand the conceptual boundaries within which the latent concept of deprivation and its dimensions operate, enriching it with the notions of variety and intensity. Furthermore, it aims to complement our understanding of the deprivation phenomenon within the segment of the older population to better assess the extent to which these people can rely upon material resources at a vulnerable stage of their lives and, ultimately, avoid the risk of poverty and social exclusion.

2 Data and Methodology

We use data from the aforementioned 2018 survey conducted by ISTAT. After selecting individuals aged 55 and over, our sample consisted of 16,515 participants. We focused on these individuals to discover the influence of varying levels of deprivation on the older population according to age and the characteristics typically linked to each specific age group.

The methodological framework employed here is based on the premise that deprivation is a multifaceted concept, encompassing various aspects of daily life, and not exclusively linked to an individual's financial circumstances. Following this reasoning and aligned with the main literature (e.g., Kim 2016), we gathered data on the material and sociodemographic indicators of deprivation rated by a set of items "that are considered by most people to be desirable or even necessary to experience an adequate quality of life" (Eurostat 2023). These include aspects such as the capacity to afford a meal with meat, chicken, and fish; maintain a sufficiently warm home; have some material goods such as a car, personal computer, telephone, and washing machine, and housing amenities such as an indoor flushing toilet. We also examine factors such as the size and condition of the home. Further, we collect data on the degradation of the neighbourhood, including pollution, crime, violence, and noise. Lastly, we include information on the ability to cover essential expenses such as housing costs.

We conduct an LCA to explore how the joint influence of the diversity and intensity of various deprivation items may lead to distinct patterns of deprivation, deriving three distinct profiles.

The LCA approach is as follows:

$$P(Y = y) = \sum_{c=0}^{c} \gamma_c \sum_{j=1}^{J} \sum_{r=1}^{R} \rho_{j,r|c}^{I(yi=rj)} \tag{1}$$

where $P(Y = y)$ is the probability of observing a set of responses, given $I(y_j = r_j)$ if the response to the variable $j = rj$, and 0 otherwise; the term γ_c indicates the probability of belonging to latent class c, while $\rho_{j,rj|c)}^{I(y_i=r_j)}$ is the likelihood of response r_j for each individual i. The key parameters here are γ, which represents the probabilities of belonging to a latent class, and ρ, which denotes the probabilities of item responses, conditional on γ. An LCA enables us to assign each sample member to the latent class that best fits their characteristics. These classes are mutually exclusive: each individual belongs to only one class. This methodology enables segmenting the sample into distinct subgroups, thereby allowing for their further empirical analyses. The core concept of LCA involves the internal creation of classes consisting of individuals with similar response patterns. We can therefore cluster individuals with comparable deprivation patterns, considering the simultaneous combination of the variety and intensity of their specific dimensions accordingly.

We begin by evaluating different models with diverse numbers of classes, ranging from a single one to five. The optimal number of classes will be determined by considering the Akaike information criterion (AIC) and the Bayes information criterion (BIC), as shown in Table 1. As the information parameters decrease as the number of classes increases, the need to preserve the parsimony and interpretability of the final model prompted us to opt for a model with fewer classes. Once the classes have been identified, a descriptive analysis is performed to investigate the levels of deprivation within the sample of individuals selected. Specifically, spider charts were employed to display multivariate data in a two-dimensional space. They are useful for comparing the deprivation patterns of older populations across different categories. Finally, probit regressions are used to identify the specific individual characteristics of older people that

Table 1. Akaike's information criterion, Bayesian information criterion, and Goodness of fit measures for LCA

Model	N	df	AIC	BIC
One class	16,494	33	452696	452950.4
Two classes	16,494	67	440473	440989.7
Three classes	16,494	101	435269.1	436047.9
Four classes	16,494	135	431821.9	432862.9
Five classes	16,494	169	430202.4	431505.5

are more likely to be correlated with the probability of experiencing deprivation across various dimensions.

3 Results

3.1 Latent Class Analysis and Descriptive Statistics

The profiles obtained through the LCA reflect the outcomes of various combinations of the variety and intensity of individual-level deprivation patterns.[1] The combinations encompass different degrees of simultaneous levels of deprivation, concerning various items. The first profile is *food deprivation*, which includes individuals with a very high conditional probability of experiencing inadequate access (rarely or never) to a sufficient quantity and quality of nutritious food, even with adequate housing conditions and an acceptable quality of life (i.e., the ability to cover essential expenses). We specifically consider access to meals that contain meat, fish, chicken, or their vegetarian equivalent (Eurostat, 2023). The second profile is *housing deprivation*. This cluster includes individuals facing significant housing challenges; specifically, higher rates of housing deprivation are linked to indicators such as neighbourhood deterioration (high levels of crime, violence, vandalism, pollution, noise, and dirt). This category also has a very high conditional probability of living in small and even unaffordable homes. Additionally, this profile also involves individuals that are unlikely to own common consumer durables, including televisions, mobile phones, personal computers, and washing machines. Finally, the third cluster is identified by *high material deprivation*. This profile includes individuals with very high conditional probabilities in all the previous deprivation indicators.

Figure 1 shows the distribution of the deprivation profiles by age. Further analyses consider distributions differentiating by gender, geographical area, employment status,

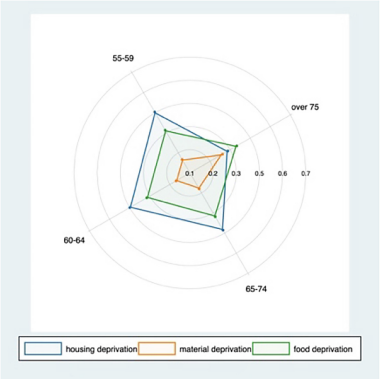


Fig. 1. Spider graph of deprivation profiles by age class

[1] Given space limitations, the parameter estimates obtained through LCA and the conditional probabilities are available upon request.

and household type.[2] These preliminary results show that food deprivation is fairly homogeneous across age groups, although it obviously records higher values compared to more severe forms of deprivation (e.g., material). Concerning housing deprivation, it has elevated values, particularly pronounced in Northern Italy, where soaring housing costs force older populations, especially those in the younger age groups, to relocate and forgo those amenities no longer considered affordable. Notably, a distinctive territorial trend also emerges in Southern Italy, where the persistent prevalence of deep-rooted economic and social precariousness in certain local contexts accentuates forms of acute material deprivation. This phenomenon, which is almost absent or mitigated in the rest of the country, records significant dimensions in comparison with Northern Italy, and maintains consistent age-related characteristics; nevertheless, housing deprivation appears to be less severe, potentially because of the prevalence of family-owned homes. Finally, there is a clear peak in material deprivation within the over-75 age group. This finding seems to reveal the worse conditions experienced by the oldest, which may be partly due to the increasing degree of social isolation in older age (Myck et al., 2020).

Addressing our first research question (RQ1), the risk of housing deprivation is higher in the 65–74 age group, while the likelihood of experiencing food deprivation is lower. There is a significant increase in high material deprivation in the over-75 age group, indicating a growing vulnerability with age (Fig. 2). Our findings on the sociodemographic characteristics of older people with a stronger association with the different dimensions of deprivation (RQ2) reveal heightened vulnerability for those individuals dependent on family maintenance or allowances, while those living together (married or cohabiting) experience less material deprivation but more housing deprivation. Education significantly influences deprivation dimensions, lowering food and material deprivation while, surprisingly, exacerbating housing deprivation. Atypical households (more than one cohabiting family unit, at least one unit with isolates) experience housing deprivation due to concentrated occupancy. In contrast, material deprivation in this household context is alleviated compared to traditional households, especially for individuals aged

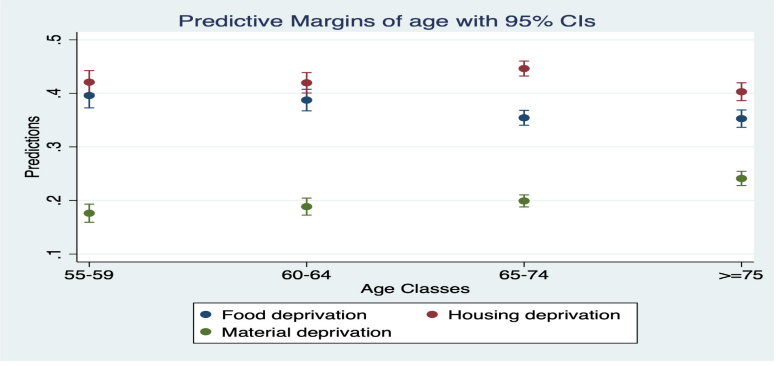


Fig. 2. Predictive Margins with 95% CI on the probability of experiencing one of the deprivation dimensions by age classes

[2] Figures available upon request.

over 75, reflecting collaborative cost reduction. Further analysis confirms the presence of a North-South deprivation divide between macro areas.

4 Conclusion

These preliminary analyses provide insights into the potentially central role of graduality in deprivation. Forms of deprivation, such as food deprivation, record a more uniform distribution within the older population. In contrast, severe forms of material deprivation are more distinctly characterised, particularly among the population over 75 living in Southern Italy, and in traditional household structures. This calls for further research in this direction, highlighting the need for nuanced studies that consider the varying degrees of deprivation across subpopulations.

By analysing deprivation among older adults, this study stresses the importance of considering the crucial role played by the combined interplay of the variety and intensity of marginal deficiencies and varying degrees of severity across different deprivation items.

References

Alkire, S., Foster, J.: Understandings and misunderstandings of multidimensional poverty measurement. J. Econ. Inequality **9**, 289–314 (2011)

Boarini, R., d'Ercole, M. M., Measures of material deprivation in OECD countries (2006)

Eurostat; Living conditions in Europe - material deprivation and economic strain, Eurostat Report (2023)

Kim, S.G.: What have we called as "poverty"? a multidimensional and longitudinal perspective. Soc. Indic. Res. **129**, 229–276 (2006)

Myck, M., Najsztub, M., Oczkowska, M.: Implications of social and material deprivation for changes in health of older people. J. Aging Health **32**(5–6), 371–383 (2020)

Navarro, C., Ayala, L.: Multidimensional housing deprivation indices with application to Spain. Appl. Econ. **40**(5), 597–611 (2006)

Panek, T.: Multidimensional approach to poverty measurement: fuzzy measures of the incidence and the depth of poverty. Stat. Trans. **11**(2), 361–379 (2010)

The Multifaced Challenge to Tackle Energy Poverty: Outlook Scenarios for EU Countries

Alfonso Carfora[1](✉) and Giuseppe Scandurra[2]

[1] University of Macerata, Macerata, Italy
alfonso.carfora@unimc.it
[2] University of Naples Parthenope, Napoli, Italy

Abstract. COVID-19 pandemic has exacerbated household energy poverty in many countries. During the latter half of 2021, as the situation gradually improved, a new threat emerged in the form of an inflationary surge in energy prices, exacerbated by the ongoing Russian-Ukrainian conflict. In this paper, we attempt to provide pragmatic projections of energy poverty in European countries over the next five years. Furthermore, the study will investigate whether nations that actively support the transition to green energy are better positioned to mitigate the adverse effects of these events on the energy poverty situation. The findings indicate that several factors contribute to energy poverty in European countries, and the short-term outlook does not look good. There is, however, a glimmer of hope for countries that prioritize and expedite their transition to green energy. A shift to renewable energy sources may contribute to a more resilient and stable energy landscape by protecting countries from the volatile nature of traditional energy markets and geopolitical conflicts. The policymakers are urged to consider sustainable energy transition not only as a means of combating climate change, but also as an essential component of economic recovery and social well-being, particularly in the context of unpredictable global events.

Keywords: energy poverty · energy transition · dynamic factors analysis

1 Introduction

The increasing incidence of energy poverty (EP) in European households is due to the effects of the COVID-19 crisis (e.g., [1]). In addition to the pandemic health crisis, the war between Russia and Ukraine has led to higher prices for energy products, exacerbating EP. The purchasing power of households, especially those most vulnerable, has been eroded by the inflation spiral triggered by the end of the acute pandemic phase and the beginning of the war between Russia and Ukraine. Recently, Carfora et al. [2] showed that the pandemic increased the share of households with heating difficulties and that the first positive effects of returning to pre-pandemic levels will not be seen until 2024. However, the proposed forecasting framework did not consider the rise in energy prices, which has impacted European households significantly. This impact is partly due to the heavy energy dependence of EU countries on foreign sources. The rise in energy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 347–353, 2025.
https://doi.org/10.1007/978-3-031-64431-3_58

prices has increased energy bills and strained households' budgets. When developing forecasting frameworks, it is essential to include the impact of energy prices to reflect the households' economic and financial conditions accurately. With the aim to better address the above-mentioned issues the European Commission (EC) decided to accelerate the transition from a predominantly fossil-based energy generation system to one more reliant on renewable sources to reduce dependence. For this reason, this paper aims to provide a valuable tool to support challenging policy decisions and identify how to reach the medium-term goals to alleviate EP. With this in mind, the paper addresses two research questions. Firstly, it aims to investigate whether countries supporting a path towards a fully green energy transition better manage price fluctuations and the adverse effects of ongoing social, political, and economic crises on energy poverty. Secondly, by providing energy poverty predictions for European countries over the next five years, it aims to investigate whether, and when European countries will be able to emerge from the deepening energy poverty that occurred as a consequence of the pandemic effects and rapid rise in energy prices.

2 Empirical Strategy

The analysis of a complex issue requires the use of quantitative methods, taking into account the complex relationships among the variables. Dynamic factor analysis (DFA) is a statistical dimension-reduction technique for time-series data that meets this need. With this method, starting from a large number of k variables, it can get an outcome of a relatively small number of m common dynamic factors. The number of estimated dynamic factors is equal to the original number of variables, but they are uncorrelated and a few factors explain most of the overall variability. In this work 15 time series have been selected to create a balanced panel dataset of 26 countries during the 2007–2020 time span. The variables were then grouped into three (k = 3) homogeneous thematic areas: i) economic and social conditions, ii) environmental degradation, and iii) housing conditions. From each thematic area a dynamic factor has been drawn. The three dynamic factors were then put into relation with an Eurostat energy poverty indicator: the % of the population reporting that they cannot keep their homes adequately warm (*enp*) and three exogenous variables, in a panel model with this structure:

$$enp_{it} = \alpha_i + \sum\nolimits_{k=1}^{3} \beta_k x_{it-1}^{k} + \boldsymbol{\gamma r}_{it} + u_{it}. \quad (1)$$

where $t = 1, \ldots, T$ is the time span (2007–2020), α_i is the individual (country), time-invariant, fixed effect, $\boldsymbol{r}$ is the matrix containing the exogenous regressors while u_{it} is the disturbance with mean equal to 0**.** Among the exogenous regressors, to take the impact of the business cycle, the gross domestic product index, provided by Eurostat, for the sample countries is used in the model. The electricity price for households (the average national price in Euro per kWh including taxes and levies for medium size household consumers) is inserted to assess the impact of the price on EP. Finally, to proxy the transition to green generation, a specific ET indicator was developed as the difference between the share of energy fed into the national energy grid generated from non-renewable sources and the share generated from renewable sources.

This fixed-effects panel specification is able to seize the relationship between energy transition and EP within EU Member States by providing useful indications on how the energy transition counteracts energy poverty and by providing forecasts of EP trends up to the 2027 upon three different scenarios:

i) Conservative Scenario: In this scenario, it is assumed that renewable energy sources will not significantly replace fossil fuels in terms of generation capacity within the next five years.
ii) Intermediate Scenario: Under this scenario, there is projected to be a gradual, linear replacement of generation capacity from fossil power plants with renewable sources over the next five years.
iii) Exponential Scenario: This scenario is akin to the intermediate one but posits a more rapid, exponential substitution of generation capacity from fossil power plants with renewable energy sources within the same five-year timeframe.

Under the scenarios developed, the expected value of EP for a generic year $t + h$. with $h = 1\ldots., n$ is conditioned on the set of information, $E(X)$, and exogenous regressors (r) that is available at time t:

$$E(Y_{t+h}) = \boldsymbol{\beta}' E(\boldsymbol{X}) + \boldsymbol{\gamma}' E(\boldsymbol{r}) + E(u_t)$$

The forecasted 2021–2027 levels of the three dynamic factors $E(X)$ are obtained based on the autoregressive path, which by construction leads to the dynamic factors:

$$\begin{aligned} E(X_{2021}) &= c + \varphi X_{2020} \\ &\vdots \\ E(X_{2027}) &= c + \varphi E(X_{2026}). \end{aligned}$$

Concerning the other regressors, the levels of GDP, expected by the International Monetary Fund World Economic Outlook (release October 2023) and the expected electricity prices on the basis of the World Bank Commodity Price Forecasts (release October 2023), have been inserted. In each of the three scenarios described above, the energy transition indicator estimated on the basis of their assumptions was included.

3 Results and Policy Implications

The estimated coefficients (Table 1) are all significant and in line with the expected signs, confirming that the proposed multidimensional latent factors and all the other exogenous regressors affect EP.

The three scenarios showed that the energy transition makes it possible to reduce the share of households reporting themselves in a condition of EP, even if it occurs with different effects depending on the assumed pathway.

Focusing on the conservative scenario (see Fig. 1), the incidence of energy poverty will decrease not due to the effect of the transition but because of the dynamics of economic growth and energy prices, as projected by the World Bank and the IMF, respectively. Countries traditionally with strong economies (such as France and Germany) or those with already low levels of energy poverty (Sweden and Finland) will experience

Table 1 Model coefficients

Description	Coefficient	std Error	p-value
Economic and social conditions dynamic factor	0.211	0.106	0.047
Environmental degradation dynamic factor	−0.638	0.127	0.000
Housing conditions dynamic factor	0.278	0.104	0.008
Gross Domestic Product index	−0.136	0.023	0.000
Households' electricity prices	29.239	10.041	0.004
Energy transition	0.247	0.049	0.000

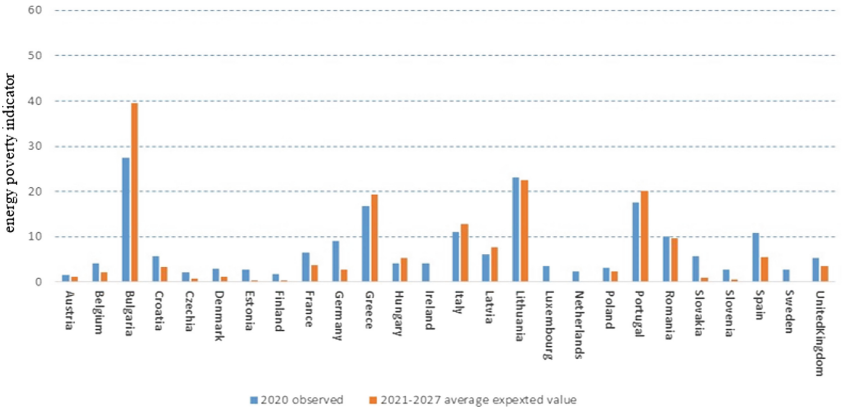


Fig. 1. Scenario 1. 2020 observed EP indicators vs average forecasted values.

a significant decrease in the percentage of households reporting difficulties in heating their homes. On the other hand, the most economically depressed countries or those characterized by high levels of energy poverty will not experience significant benefits.

The effects of the transition will be more pronounced in the other two scenarios (see. Figs. 2 and 3), where, while maintaining the characteristics of the first scenario, a substantial reduction in the incidence of energy poverty will be observed. In general, it is observed that the coming years should see an average reduction in the incidence of energy poverty in European countries, although there are clear signs that it will still lead to an increase in some nations. The energy transition, including the reduction in electricity costs (e.g., [3, 4]), will bring the expected benefits to families in need, which an anticipated phase of economic expansion will amplify. As a consequence of the mentioned conditions, international, European, and national authorities should recognize that country-specific interventions are needed to prevent further increases in inequality as EP increases. It is acknowledged that actions to tackle EP may vary depending on the context and peculiarities of each country. As demonstrated, decisive action toward energy transition is certainly helpful in reducing the impact of EP. Clearly, a linear increase is not enough, but substantial investment is needed that will lead to exponential growth in

the amount of electricity produced from renewables that gradually replaces that portion generated from traditional sources. Moreover, investments in the energy transition will, over the years, lead to lower electricity rates (e.g., [5]).

In the meantime, it would be helpful to consolidate all the diverse electricity and gas benefits into a singular tool linked to each household's individualized risk of EP. This is a highly ambitious target, considering the variations in measurement criteria and the definition of EP across the European Union. Further efforts should be devoted to improving vulnerable households' living and housing conditions [6]. In a historical stage marked by high energy prices, policy authorities should envision (or strengthen if already in place) energy subsidy programs to help vulnerable households improve the condition of their homes by reducing energy consumption and improving efficiency. Investing in energy efficiency and building upgrading projects can reduce households' energy consumption and, consequently, their expenditures. Proposals for reducing energy consumption should consider that a large percentage of households in the EU live in houses with inadequate insulation. This situation may be exacerbated by climate change. Previous studies, such as the one conducted by Damigos et al. [7], have shown that low-income households tend to focus on short-term rather than long-term outcomes, making them prone to make limited decisions (the so-called "discounting gap"). The "Recovery Plan" might play a crucial role in this perspective. Following the 2020 EC recommendations, these funds are expected to be allocated to support households in lowering energy expenses, improving quality of life. Our findings emphasize that the faster the energy transition, the greater the action to counter energy poverty can be.

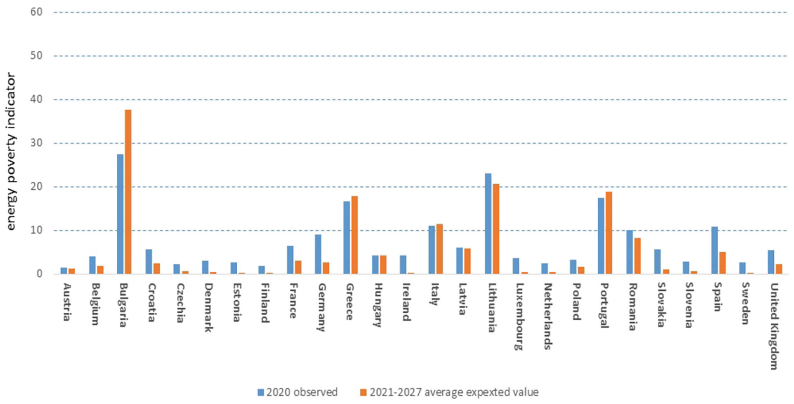


Fig. 2. Scenario 2. 2020 observed EP indicators vs average forecasted values.

Moreover, as further policy implication, encouraging the adoption of renewable energy can help reduce long-term energy costs and improve household energy resilience. Indeed, countries less affected by energy poverty are also the ones in which the share of energy generated from renewable sources fed into the grid predominates over that generated from fossil fuels. The acceleration of the energy crisis due to the recent war, which started after the pandemic period and the resulting severe economic consequences, shows a not encouraging outlook for countries. However, the current challenges can be

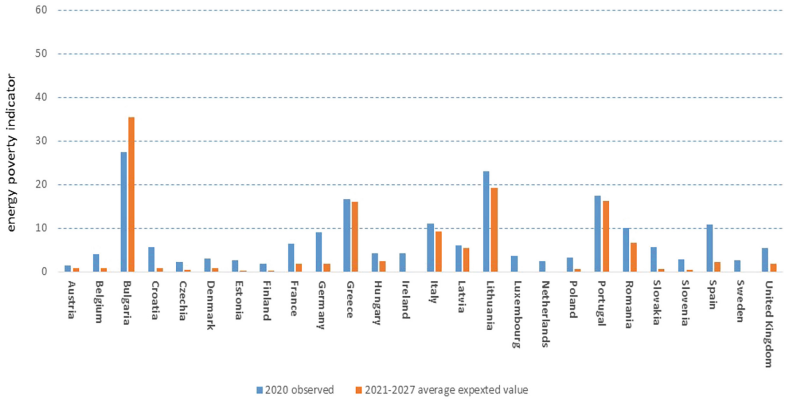


Fig. 3. Scenario 3. 2020 observed EP indicators vs average forecasted values.

turned into opportunities for renewable energy over the next few years. An example is offered by taking advantage of benefits from the digitization of work and other daily activities, as it could reduce traveling and the consumption of fossil energy sources [8]. A recommendation urges to promote the collaboration among stakeholders (policymakers, nongovernmental organizations, research institutions, energy suppliers, and others) can help develop integrated policies and customizing solutions to address energy poverty. The synergy can lead to greater effectiveness in identifying needs and implementing targeted interventions that take into account unique country's characteristics, such as different climatic and socioeconomic conditions.

References

1. Manjon, M.J., Merino, A., Cairns, I.: Business as not usual: a systematic literature review of social entrepreneurship, social innovation, and energy poverty to accelerate the just energy transition. Energy Res. Soc. Sci. **90**, 102624 (2022)
2. Carfora, A., Scandurra, G., Thomas, A.: Forecasting the COVID-19 effects on energy poverty across EU member states. Energy Policy **161**, 112597 (2022). https://doi.org/10.1016/j.enpol.2021.112597
3. Kammer, A.: Europe Must Address a Toxic Mix of High Inflation and Flagging Growth. IMF blog (2022). https://www.imf.org/en/Blogs/Articles/2022/10/23/europe-must-address-a-toxic-mix-of-high-inflation-and-flagging-growth
4. Aghahosseini, A., et al.: Energy system transition pathways to meet the global electricity demand for ambitious climate targets and cost competitiveness. Appl. Energy **331**, 120401 (2023). https://doi.org/10.1016/j.apenergy.2022.120401
5. Cevik, S., Ninomiya, K.: Chasing the Sun and Catching the Wind: Energy Transition and Electricity Prices in Europe. IMF Working Paper, WP/22/220, IMF (2022)
6. Fell, M.J., et al.: Validity of energy social research during and after COVID-19: challenges, considerations, and responses. Energy Res. Soc. Sci. **68**, 101646 (2020)

7. Damigos, D., Kaliampakou, C., Balaskas, A., Papada, L.: Does energy poverty affect energy efficiency investment decisions? first evidence from a stated choice experiment. Energies **14**, 1698 (2021)
8. Abu-Rayash, A., Dincer, I.: Analysis of the electricity demand trends amidst the COVID-19 coronavirus pandemic. Energy Res. Soc. Sci. **68**, 101682 (2020)



Bayesian Models for Registration of Functional Data

Jacopo Gardella[1], Rossella Carone[1], Simone Ciardulli[1], Nicola Francescon[1], Matteo Freddi[1], Elisa Garlanda[1], Alessandro Casa[2], and Alessia Pini[3(✉)]

[1] Politecnico di Milano, MSc Mathematical Engineering, Milan, Italy
[2] Department of Economics, University of Bergamo, Bergamo, Italy
[3] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy
alessia.pini@unicatt.it

Abstract. Studies involving functional data frequently require curve registration, i.e. the alignment of salient features in the temporal domain, as a preliminary step before the application of inferential techniques. Through this procedure, phase variability becomes negligible, thus allowing to focus solely on amplitude variability. This work introduces a Bayesian model for curve alignment and applies it to a biomechanical dataset with three groups of patients.

Keywords: Functional data · Bayesian warping · B-splines · Markov chain Monte Carlo

1 Introduction

Functional data analysis deals with the examination of random curves, known as functional data. Such data often exhibit a common shape with dissimilarities in amplitude and phase, as described by Ramsay, Silverman (2005) [1]. Subsequently, difficulties come to light when determining if there are any noteworthy differences among groups of subjects or in defining the shape of the observed process. Curve alignment is thus a preliminary task that is often necessary to make statistical inference.

The present work aims to provide a model for curve registration by finding warping functions that synchronize all the curves. Such an analysis is conducted in a Bayesian framework and relies on some previous work on the matter.

An earlier research involving the same data has been conducted by Abramowicz et al. (2018) [2]. The data were aligned using a landmark-based registration of the curves. However, landmark registration presents potential drawbacks: the exact location of the feature could be unclear for specific curves, and since the alignment focuses only on single points, discrepancies in other areas may be overlooked. Moreover, the statistical analysis conducted in this research did not consider the uncertainty associated with the curve alignment.

The current work proposes a model-based Bayesian approach for curve alignment, starting from a work presented by Telesca and Inoue (2008) [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 354–359, 2025.
https://doi.org/10.1007/978-3-031-64431-3_59

2 Model

Let $y_i(t_{ij})$ denote the observed knee flexion angle of the i-th patient at time t_{ij}, with $i = 1, \ldots, N$ and $t_{ij} \in T = [0, 1]$, $j \in \{1, \ldots, n_i\}$, where n_i denotes the number of time instants for which y_i has been collected. Following [3], data are assumed to share a common shape function

$$m(t; \boldsymbol{\beta}) = \mathcal{B}_m(t)^T \boldsymbol{\beta}$$

where $\mathcal{B}_m(t)$ denotes a set of cubic B-spline basis functions with f internal knots. Therefore, $\boldsymbol{\beta} \in \mathbb{R}^p$, where $p = f + 4$.
The individual shape function of subject i is then constructed by means of a linear transformation of $m(t; \boldsymbol{\beta})$:

$$m_i(t) = c_i + a_i m(t; \boldsymbol{\beta})$$

where the coefficients c_i and a_i are subject-specific. The final observation of subject i is assumed to be obtained by the composition of the individual shape function $m_i(t)$ and an individual warping function $h_i(t)$. Individual warping functions are again built using B-spline basis functions with subject-specific coefficients:

$$h_i(t) = \mathcal{B}_h(t)^T \boldsymbol{\phi}_i$$

once again, having fixed d internal knots. Then $\boldsymbol{\phi}_i \in \mathbb{R}^q$, $i = 1, \ldots, N$, with $q = d + 4$. Finally, the observed curves y_i can be modelled as

$$y_i(t) = m_i[h_i(t)] + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

To retrieve the aligned curves y_i^*, it is possible to apply the inverse warping functions to the original curves

$$y_i^*(t) = y_i[h_i^{-1}(t)]$$

To let the problem be consistent with the physical constraints, the following assumptions are required:

- Time is constrained to be in the interval $[0, 1]$:

$$h_i(t_{i1}) = h_i(0) = 0, \qquad h_i(t_{in_i}) = h_i(1) = 1, \qquad i = 1, \ldots, N \tag{1}$$

  To satisfy this constraint it is sufficient to impose $\phi_{i1} = 0$, $\phi_{iq} = 1$, $i = 1, \ldots, N$
- Time transformations should be monotone in order not to switch the ordering of the observations:

$$h_i(t_{i1}) < h_i(t_{i2}) < \ldots < h_i(t_{in_i}) \qquad i = 1, \ldots, N \tag{2}$$

  As stated by De Boor (2001) [7], to meet the assumptions (1) and (2), it is sufficient that

$$0 = \phi_{i1} < \phi_{i2} < \ldots < \phi_{iq} = 1 \qquad i = 1, \ldots, N \tag{3}$$

The prior distribution is directly derived from [3], but a modification on the part concerning coefficients $\boldsymbol{\phi}_i$ is introduced, inspired by the methodology proposed by Montagna, Orani & Argiento (2019) [4]. The new prior distribution intrinsically guarantees the monotonicity constraint (3) in model formulation by introducing non-negative vectors $\boldsymbol{\xi}_i$, so that each component of $\boldsymbol{\phi}_i$ is obtained as the cumulative sum of the components of the corresponding $\boldsymbol{\xi}_i$. The resulting prior distribution can be consequently expressed as follows:

$$\phi_{ij} = \frac{\sum_{k=1}^{j} \xi_{ik}}{T_i} \qquad i = 1, \ldots, N \qquad j = 1, \ldots, q$$

where

$$\xi_{ij} \overset{iid}{\sim} gamma(a, b) \qquad i = 1, \ldots, N, \qquad j = 2, \ldots, q$$

$$\xi_{i1} = 0, \qquad T_i = \sum_{k=1}^{q} \xi_{ik} \qquad i = 1, \ldots, N$$

2.1 Implementation

An MCMC method is employed to sample from the full conditional distributions of the parameters. Whilst most parameters admit a closed form full conditional distribution, an exception arises with the parameters $\boldsymbol{\phi}_i$, whose update is carried out indirectly through the update of the $\boldsymbol{\xi}_i$ variables, performed separately on each component of the vector.
The Metropolis-Hastings step works as follows:

1. ξ_{i1}^{new} is set to zero, and consequently

$$\phi_{i1}^{new} = \frac{\xi_{i1}^{new}}{T_i} = 0$$

2. For $j = 2, \ldots, q$ the algorithm proposes a new value $\xi_{ij}^{new} \sim gamma(a_{prop}, b_{prop})$.

The parameters of the proposal distribution vary across iterations. Adaptive Metropolis-Hastings is employed in order to automatize the manual tuning and possibly speed up the convergence. Specifically, the hyperparameters a_{prop} and b_{prop} are updated at each iteration in such a way that $\mathbb{E}[\xi_{ij}^{new}] = \xi_{ij}^{old}$, where ξ_{ij}^{old} is the previous value of the chain, according to the relations:

$$a_{prop} = \frac{1}{v^2}$$

$$b_{prop} = \frac{a_{prop}}{\xi_{ij}^{old}}$$

The variation coefficient $v = 0.3$ is set to guarantee that the MCMC explores a sufficient range of values.

As a direct consequence, all the elements of the proposed vector $\boldsymbol{\phi}_i^{new}$ are modified accordingly as follows:

$$\phi_{ih}^{new} = \begin{cases} \frac{\sum_{k=1}^{h} \xi_{ik}^{prev}}{T_i^{new}} & \text{if } h = 1, \ldots, j-1 \\ \frac{\sum_{k=1}^{j-1} \xi_{ik}^{prev} + \xi_{ij}^{new}}{T_i^{new}} & \text{if } h = j \\ \frac{\sum_{k=1 k \neq j}^{j-1} \xi_{ik}^{prev} + \xi_{ij}^{new} + \sum_{k=j+1 k \neq j}^{h} \xi_{ik}^{old}}{T_i^{new}} & \text{if } h = j+1, \ldots, q \end{cases}$$

where ξ_{ij}^{prev} are the values of the components determined until the current step and can be different from ξ_{ij}^{old}. Indeed, the former may be either the result of an acceptance of the current step of the Metropolis-Hastings algorithm or come from a previous iteration of the MCMC as the corresponding new proposal point was rejected.

Remarkably, the acceptance of a new proposal point modifies all the components of the vector $\boldsymbol{\phi}_i$, both the ones that the algorithms have already updated and those that have not yet been updated.

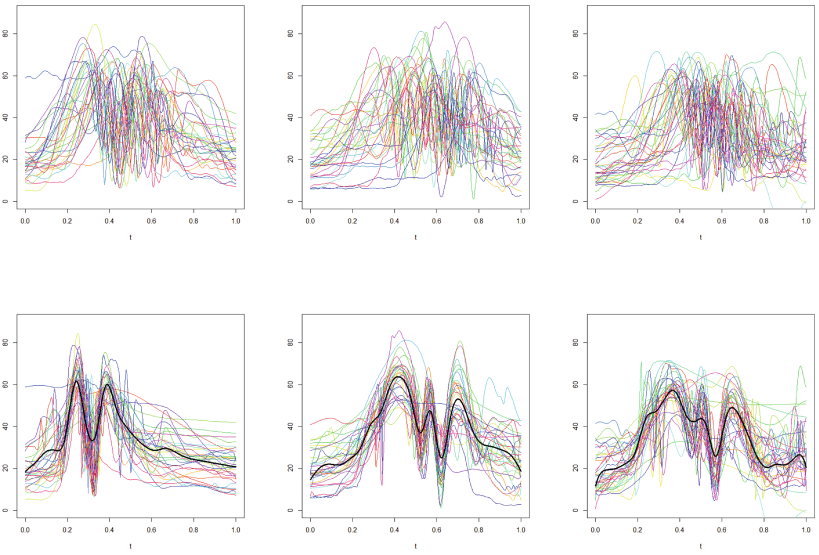
3 Results

This work analyses a follow-up study of subjects who faced anterior cruciate ligament (ACL) injury. More in detail, the observed curves describe the knee flexion angle during a one-leg hop of three groups of patients - undergone surgery, undergone physiotherapy and healthy (control group) - comprehending 32, 36 and 32 patients, respectively. Data are provided by the U-motion Laboratory, Umeå University, Sweden, and their detailed description can be found in [8]. The goal is to detect differences between the three groups. Since the length in time of a jump is not constant between patients, a preliminary step involving curve registration is needed.

The inference is based on 15.000 samples from the posterior distribution obtained after discarding the initial 50.000 MCMC burn-in iterations. The model has been applied separately to the data of the three groups in order to appreciate the group's differences. The aligned curves are displayed in Fig. 1.

The functional behavior appears different among the groups. While the common shape function of the surgery patients presents the three expected peaks [2], the other groups display only two peaks. Therefore, a further investigation of the different behaviors of the three groups should be taken into consideration. For the physiotherapy group, instead, prior knowledge [2] shows that these patients tend to have knee stability issues mainly in the landing phase, during which the knee is subject to major strain. Indeed, the common shape function turns out to be quite noisy.

These three alignments look fairly reasonable: although the model is unable to effectively register all the curves, it aligns most of them, and it captures the general pattern. Healthy patients seem the most troublesome group, since the



(a) Healthy (b) Surgery (c) Physiotherapy

Fig. 1. Above the observed curves of the three groups, below their respective alignment and in black the group-specific common shape functions

model misses the third peak and presents an undesired behavior between the two peaks.

Overall, the comparison made on the different groups results to be coherent with both the prior scientific knowledge on the problem and the (non-Bayesian) approach by Abramowicz et al. [2].

4 Conclusions and Future Developments

In this work a modification of the model by Telesca has been proposed, in particular a new prior distribution for the ϕ_i coefficients has been used in order to directly guarantee the monotonicity assumption, following the approach proposed by Montagna, Orani & Argiento (2019) [4]. The alignment of these three curves appears generally good, capturing the overall pattern. However, the possible presence of some outliers and noisy observations may negatively impact the result. Therefore, an outlier detection analysis of the curves could be conducted to spot and remove some possible anomalous observations and increase the robustness of the procedure.

Furthermore, the model proposed by Telesca and Inoue (2008) [3] has intrinsic limitations in terms of flexibility. Specifically, the common shape function, which can only be shifted and dilated, may not optimally capture individual variations among patients. To relax such a restrictive approach and introduce greater flexibility into the model, an extension based on the work by Montagna

et al. (2012) [5] can be proposed, in which a Bayesian latent factor model is used with an adaptive Gibbs sampler approach as described in Dunson, Bhattacharya (2011) [6]. This model would better adapt to specific patient characteristics and could be more suitable for noisy data.

Finally, since the patients are divided in three groups, another possible future development could consider to fit a hierarchical model by including the group as a categorical feature.

Acknowledgements. We wish to acknowledge the U-motion laboratory of Umeå University and Charlotte K. Hager for providing the knee flexion data.

References

1. Ramsay, J.O., Silverman, B.W. *Functional Data Analysis*. Springer (2005)
2. Abramowicz, K., Häger, C.K., Pini, A., Schelin, L., Sjöstedt de Luna, S., Vantini, S. Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. Scandinavian J. Stat. **45**(4), 1036–1061 (2018)
3. Telesca, D., Inoue, L.Y.T.: Bayesian Hierarchical Curve Registration. J. Am. Stat. Assoc. **103**(481), 328–339 (2008)
4. Montagna, S., Orani, V., Argiento, R.: Bayesian isotonic logistic regression via constrained splines: an application to estimating the serve advantage in professional tennis. Stat. Methods Appl. **30**, 573–604 (2019)
5. Montagna, S., Tokdar, S.T., Neelon, B., Dunson, D.B.: Bayesian latent factor regression for functional and longitudinal data. Biometrics **68**(4), 1064–1073 (2012)
6. Bhattacharya, A., Dunson, D.B.: Sparse Bayesian infinite factor models. Biometrika **98**(2), 291–306 (2011)
7. De Boor, C.: A practical guide to splines. Springer-Verlag (2001)
8. Hèbert-Losier, K., et al.: One-leg hop kinematics 20 years following anterior cruciate ligament rupture: data revisited using functional data analysis. Clin. Biomech. **30**(10), 1153–1161 (2015)



Regionalization Methods for Compound Extremes Based on the Wasserstein Distance

Regina Castrovilli[1], Fabrizio Durante[1], Daniela Gallo[1,2], and Gianfausto Salvadori[1(✉)]

[1] Department of Mathematics and Physics "Ennio De Giorgi", University of Salento (Italy), Lecce, Italy
{regina.castrovilli,fabrizio.durante,daniela.gallo, gianfausto.salvadori}@unisalento.it

[2] Institute for High Performance Computing and Networking, Rende (Italy), Arcavacata, Italy
daniela.gallo@icar.cnr.it

Abstract. A clustering algorithm is presented in order to group time series of maxima for pair of random variables recorded at different stations. The method is based on two kinds of dissimilarity measure. The first one involves the distance of the marginal distributions; the second one involves the distance between the respective copulas. In both cases, a Wasserstein metric is used. For the estimation, we rely on tools for extreme-value distributions and copulas.

Keywords: Clustering · Copula · Extreme Values · Unsupervised learning

1 Introduction

Statistical modeling of climate extremes such as heatwaves, droughts, and floods is of utmost interest since it can help to mitigate potentially severe impacts on society [6]. In particular, many large impacts are caused by Compound Events, i.e. the concurrent occurrence of multiple and possibly interdependent hazards, whose impact can be large as compared to univariate extremes [4]. In such cases, univariate techniques accounting for the effect of each hazard separately may underestimate the true overall risk [17].

One main line of research in statistical climatology is the *regionalization*, i.e. the identification of sub-regions characterized by a "similar" behavior of the variables of interest [13]: this is of fundamental importance for two main reasons [8]. Firstly, this allows climate scientists to gain a deeper knowledge of the phenomenon under consideration, as well as its different spatial effects. Secondly, determining regions that can be roughly considered statistically similar (in terms of their distributional behavior) to a given location of interest, can

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 360–365, 2025.
https://doi.org/10.1007/978-3-031-64431-3_60

improve the estimation of specific quantities at this latter site when records are scarce.

Many clustering algorithms have been proposed to summarize spatial complex climatological datasets, and to capture some features of their underlying hidden patterns, also by focusing on extremes [1,3,12]. Here, the main issue is to develop a clustering algorithm that takes into account the climatological similarity of a multivariate phenomenon (in the sense of Compound Events) observed at different sites. Specifically, in the framework of block-maxima methodology [15], we propose two dissimilarity measures that are related to (a) the univariate distribution of extremes, and (b) the copula associated with these maxima. In both cases, we rely on the Wasserstein distance (see, e.g., [16]), which allows for a meaningful comparison between distributions also without density and, in some applications, does not lead to counter-intuitive clusters (see, e.g., [2,10]).

2 Methodology

Consider a set of n gauge stations over a given geographical area. At each station i, we collect time series of observations $\{(x_t^i, y_t^i)\}_{t=1,\dots,T}$ about two (or more) attributes (X_i, Y_i). Such observations are related to maxima (or minima) of the features of interest over a given time period. Note that minima can always be transformed into maxima (and vice versa) simply by changing the sign of the variable.

Our aim is to provide a procedure to cluster the stations according to two types of information: (a) the (univariate) distribution of one attribute; (b) the copula associated with both the attributes at each site. The clustering procedure embraces the issue of Hierarchical Clustering as formalized in [7]. Namely, we define a pairwise dissimilarity between all the stations that depends on their marginal (respectively, joint) probability law.

As for the dissimilarity measure, we will leverage on the 2-Wasserstein Distance between probability measures [11]. Given a domain $\Omega \subseteq \mathbb{R}^d$ (assumed to be compact and convex for simplicity), we define the space $\mathcal{W}_2(\Omega)$ as the space of probability measures on Ω with finite moments of order 2, endowed with the distance d_W defined by

$$d_W(\mu_X, \mu_Y) = \left(\inf_{\gamma \in \Gamma(\mu_X, \mu_Y)} \int_{\Omega \times \Omega} |x - y|^2 d\gamma(x, y) \right)^{1/2}, \tag{1}$$

where $\Gamma(\mu_X, \mu_Y)$ denotes the collection of all joint measures on $\Omega \times \Omega$ whose marginals are μ_X and μ_Y, respectively.

In particular, in order to measure the dissimilarity between the variables X_i and X_j, $i \neq j$, we consider the formula

$$d_W^2(X_i, X_j) = \int_0^1 |F_i^{-1}(u) - F_j^{-1}(u)|^2 du, \tag{2}$$

where F_i^{-1} and F_j^{-1} are the quantile functions associated with X_i and X_j, respectively, and similar formulas for the Y's. The estimation of such a distance is carried out via two steps.

1. For all $i = 1, \ldots, n$, a univariate GEV law F is fitted to the time series $(x_t^i)_{t=1,\ldots,T}$; here,

$$F(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi \frac{x-\mu}{\sigma}\right]^{-1/\xi} \right\}, \tag{3}$$

with $1 + \epsilon \frac{x-\mu}{\sigma} > 0$, where $(\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R})$ are, respectively, position, scale, and shape parameters. In turn, the quantile function Q associated with F (i.e., $Q = F^{-1}$) is given by

$$Q(p; \mu, \sigma, \xi) = \mu + \frac{\sigma}{\xi}\left[(-\log(p))^{-\xi} - 1\right], \tag{4}$$

where $p \in (0, 1)$ is a probability level. Since the considered time series are related to maxima of variables, the GEV distribution provides a theoretical well-grounded choice [14].
2. A Maximum Likelihood estimate $\widehat{\theta}$ of the GEV parameters of F is calculated.
3. Finally, $diss(X_i, X_j) = d_W(X_i, X_j)$ is estimated via numerical integration, exploiting the quantile function Q shown in the previous step.

In the same vein of thinking, in order to estimate the similarity between the copulas C_i of (X_i, Y_i) and C_j of (X_j, Y_j), we consider the class of bivariate Extreme Value (EV) copulas [5]. Specifically, to each EV copula C we can associate a Kendall distribution function K_C given by, for $u \in [0, 1]$, $K_C(u) = u - (1 - \tau)u\log(u)$, where $\tau = \tau(C)$ is the Kendall's measure of association related to C [9]. If (U, V) is a random pair distributed according to C, then K_C is the distribution function of the random variable $C(U, V)$. Hence, we define the dissimilarity between C_i and C_j as

$$diss(C_i, C_j) = d_W(K_{C_i}, K_{C_j}), \tag{5}$$

i.e. in terms of the dissimilarity between the related Kendall distribution functions. Easy calculations show that such a dissimilarity is related to the distance $|\tau(C_i) - \tau(C_j)|$ between the related Kendall's correlation coefficients. Thus, the estimation of such a dissimilarity can be carried out via a standard estimation of the corresponding Kendall τ's.

3 Illustration: Analysis of Heat Waves

In order to show the proposed methodology we consider daily maxima temperature levels (in Celsius) and minima relative humidity in 22 municipalities located in Apulia region (Southern Italy) collected by ARPA Puglia from January 1st, 2019 to December 31th, 2022. Therefore, we have $n = 22$ time series and $T = 1461$ observations. In order to consider pairs of maxima, the sign of the relative humidity has been reversed.

The investigation has been carried out by adopting a copula-based approach to multivariate time series. Concerning the univariate time series, as is typical

in geophysical investigations, first we remove seasonal effects by subtracting a deterministic seasonal component [1,3]. Then, we standardize the observations in order to get values that are physically comparable among different sites. Practically, each time series is normalized by first subtracting its sample mean, and then by dividing by its sample standard deviation.

Finally, a classical Agglomerative Hierarchical Clustering (complete linkage) [7] is applied to the residuals: here, the dissimilarity measures are those outlined in the previous sections.

Specifically, in Fig. 1, we present the dendrogram obtained for the temperature and the humidity variables, respectively. While in Fig. 2, we show the dendrogram obtained from the associated copulas. All these graphs may provide

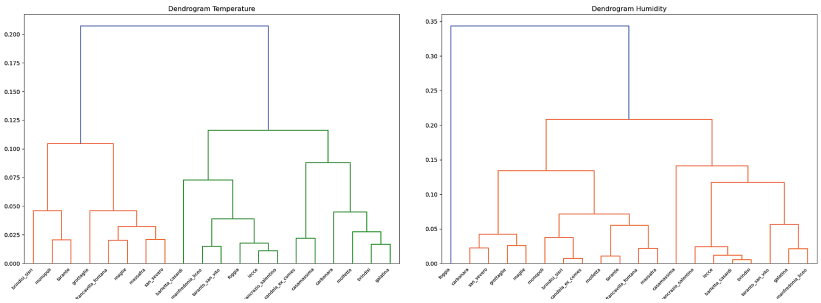


Fig. 1. Dendrogram obtained from hierarchical agglomerative clustering (complete linkage) applied to temperature (left) and relative humidity (right) time series with dissimilarity defined via (2).

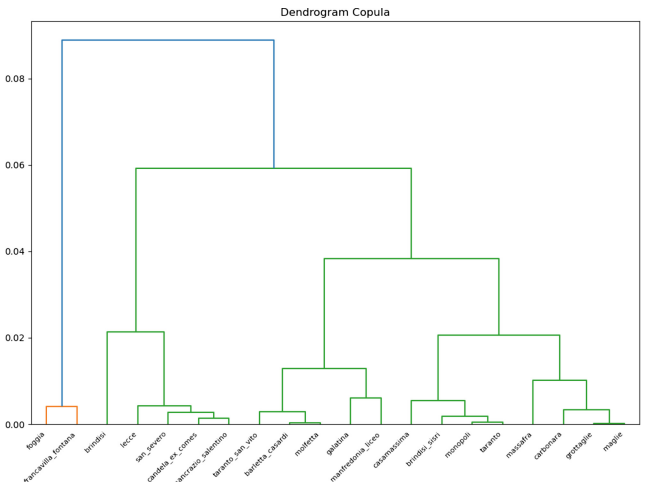


Fig. 2. Dendrogram obtained from hierarchical agglomerative clustering (complete linkage) applied to the copula of temperature and relative humidity time series with dissimilarity defined via (5)

useful information to understand the similarities among these variables across different stations and may help describing high temperature/low humidity phenomena, also called heat waves.

Acknowledgments. FD and GS have been supported by MIUR-PRIN 2017, Project "Stochastic Models for Complex Systems" (No. 2017JFFHSH) and by MUR-PRIN 2022 PNRR, Project "Stochastic Modeling of Compound Events" (No. P2022KZJTZ) funded by European Union – Next Generation EU. The work of RC, FD, and GS has been carried out with the partial financial support from ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – Next Generation EU (CUP F83C22000740001).

DG has been supported by MUR-PRIN 2022, Project "WHAM!: Watermarking Hazards and novel perspectives in Adversarial Machine" (Code: 2022ZZX57L, CUP B53D23013340006) funded by European Union – Next Generation EU.

References

1. Bador, M., Naveau, P., Gilleland, E., Castellà, M., Arivelo, T.: Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. Weather Clim. Extremes **9**, 17–24 (2015)
2. Benevento, A., Durante, F.: Wasserstein dissimilarity for copula-based clustering of time series with spatial information. Mathematics **12**(1), 67 (2024)
3. Bernard, E., Naveau, P., Vrac, M., Mestre, O.: Clustering of maxima: spatial dependencies among heavy rainfall in France. J. Clim. **26**(20), 7929–7937 (2013)
4. Bevacqua, E., et al.: Guidelines for studying diverse types of compound weather and climate events. Earth's Future **9**(11), e2021EF002,340 (2021)
5. Durante, F., Sempi, C.: Principles of Copula Theory. CRC Press, Boca Raton, FL (2016)
6. Field, C.B.: Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press (2012)
7. Fuchs, S., Di Lascio, F.M.L., Durante, F.: Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. Comput. Statist. Data Anal. p. 107201 (2021)
8. Grimaldi, S., Petroselli, A., Salvadori, G., De Michele, C.: Catchment compatibility via copulas: a non-parametric study of the dependence structures of hydrological responses. Adv. Water Resour. **90**, 116–133 (2016)
9. Gudendorf, G., Segers, J.: Extreme-value copulas. In: Jaworski, P., Durante, F., Härdle, W.K., Rychlik, T. (eds.) Copula Theory and its Applications. Lecture Notes in Statistics - Proceedings, vol. 198, pp. 127–145. Springer, Berlin Heidelberg (2010)
10. Marti, G., Andler, S., Nielsen, F., Donnat, P.: Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series. In: 2016 IEEE statistical signal processing workshop (SSP), pp. 1–5. IEEE (2016)
11. Panaretos, V.M., Zemel, Y.: Statistical aspects of Wasserstein distances. Annu. Rev. Stat. Appl. **6**, 405–431 (2019)
12. Pappadà, R., Durante, F., Salvadori, G., De Michele, C.: Clustering of concurrent flood risks via hazard scenarios. Spat. Stat. **23**, 124–142 (2018)

13. Parajka, J., Merz, R., Blöschl, G.: A comparison of regionalisation methods for catchment model parameters. Hydrol. Earth Syst. Sci. **9**(3), 157–171 (2005)
14. Salvadori, G., De Michele, C., Kottegoda, N., Rosso, R.: Extremes in nature: an approach using copulas, vol. 56. Springer Science & Business Media (2007)
15. Salvadori, G., Durante, F., De Michele, C., Bernardi, M., Petrella, L.: A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. Water Resour. Res. **52**(5), 3701–3721 (2016)
16. Santambrogio, F.: Optimal transport for applied mathematicians. Birkäuser, NY **55**(58–63), 94 (2015)
17. Vignotto, E., Engelke, S., Zscheischler, J.: Clustering bivariate dependencies of compound precipitation and wind extremes over great Britain and Oreland. Weather Climate Extremes **32**, 100,318 (2021)



Efficient Parametric Tests in Semiparametric Regression with Differential Regularization

Michele Cavazzutti[1(✉)], Eleonora Arnone[2], Federico Ferraccioli[3], Livio Finos[3], and Laura M. Sangalli[1]

[1] MOX laboratory, Politecnico di Milano,Milan, Italy
michele.cavazzutti@polimi.it
[2] Università degli Studi di Torino,Turin, Italy
[3] Università degli Studi di Padova,Padua, Italy

Abstract. In this work we consider the problem of making inference on the nonparametric component within a semiparametric regression model with differential regularization. The parametric inference methods so far introduced in the literature perform poorly, due to the variance misspecification induced by the penalization term. Nonparametric inference procedures may instead be excessively computationally demanding. We hereby propose two new parametric approaches, that are robust to the effect of the penalization, while retaining a reduced computational cost. The first method relies on an appropriate undersmoothing strategy, combined with a bootstrap approach. The second one leverages instead on the asymptotic properties of the scores of the model. The resulting tests have better control of Type-I error, with respect to the existing alternatives, and reduced computational cost. We apply the novel approaches to the study chlorophyll-a concentrations in the Mediterranean sea.

Keywords: Smoothing · roughness penalty · semiparametric regression · bootstrap

1 Introduction

We address the problem of making inference on the nonparametric component of a semiparametric regression model with differential regularization. Semiparametric regression with roughness penalties is a wide class of models, where the response variable is modeled as the sum of the contribution of a parametric and a nonparametric term.

Various approaches to make inference on the nonparametric term have been attempted in the literature of nonparametric and semiparametric regression models. See, e.g., [4,5,9,14] for some examples of inference in these contexts. In this work, we are mainly interested, but not restricted, to the class of Spatial Regression with Partial Differential Equation regularization (SR-PDE) models,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 366–372, 2025.
https://doi.org/10.1007/978-3-031-64431-3_61

first presented in [11] and reviewed in [10]. Recently [3,6,7] have studied some inference approaches for these models. Regarding the nonparametric component, we can distinguish among parametric and noparametric inference methods. Approaches of the first kind usually have a Wald-like form, and rely on the asymptotic properties of the estimator of the model. These methods are flexible in the type of hypotheses that can be tested, and in general are computationally cheap. However, the Wald-like tests are heavily affected by the variance misspecification and the bias induced by the smoothing penalty, and in practice have poor control of Type-I error in the finite sample scenarios. [3] first explores some new nonparametric approaches for the nonparametric term of SR-PDE. Such tests, named Sign-Flip tests, rely on the innovative resampling procedure described in [8]. These tests have good control of the Type-I error and high power, but in turn are computationally expensive. In particular, in the large sample scenarios, the computation of confidence intervals based on the Sign-Flip tests becomes quickly unfeasible. This is a major drawback, as confidence intervals are fundamental for the understanding and the visualization of a functional field.

In literature, the variance misspecification induced by the regularization of the model is frequently addressed by means of undersmoothing approaches (see, e.g., [12,14] for some examples in the general semiparametric regression setting). Following this approach, we here propose a bootstrap procedure to correct the estimate of the variance of a Wald-like statistic based on the estimator of the model. In addition, we also develop a novel test for the nonparametric component of SR-PDE, based on the asymptotic properties of the score equations of the model. We employ the proposed methods to study the concentration of Chlorophyll-a in Mediterranean sea, illustrated in Fig. 1.

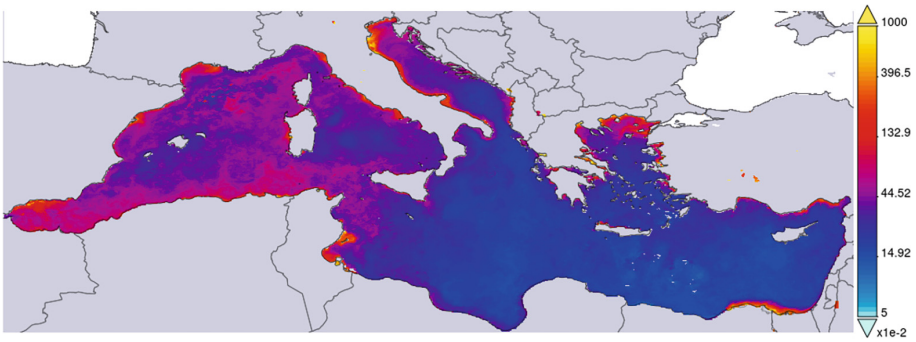


Fig. 1. Chlorophyll-a concentration $[mg \cdot m^{-3}]$ represented with logarithmic color scale in Mediterranean sea in February 2019, reproduced using NASA application GIOVANNI.

2 Spatial Regression with Partial Differential Equation Regularization

We here recall the definition of a Spatial Regression with Partial Differential Equation regularization (SR-PDE) model. Consider a domain $\Omega \subset \mathbb{R}^2$ of limited measure. Let $y \in \mathbb{R}$ be a variable of interest, and $\mathbf{x} \in \mathbb{R}^q$ a set of covariates. Let $\{\mathrm{y}_i\}_{i=1,\dots,n}$ and $\{\mathbf{x}_i\}_{i=1,\dots,n}$ be the observations of y and $\mathbf{x}$ on a set of n locations $\{\mathrm{p}_i\}_{i=1,\dots,n}$ distributed over Ω. We model the response variable such that, for an unknown vector $\boldsymbol{\beta} \in \mathbb{R}^q$ and function $f : \Omega \longrightarrow \mathbb{R}$,

$$\mathrm{y}_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(\mathbf{p}_i) + \epsilon_i \tag{1}$$

for $i = 1, ..., n$, where $\{\epsilon_i\}_i$ is a set of i.i.d. errors with zero mean and finite variance σ^2. The unknowns $\boldsymbol{\beta}$ and f are estimated by minimization of

$$J(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \left(\mathrm{y}_i - \mathbf{x}_i^\top \boldsymbol{\beta} - f(\mathbf{p}_i)\right)^2 + \lambda \int_\Omega (Lf)^2, \tag{2}$$

where $Lf = 0$ is a linear elliptic PDE that describes the available expert knowledge on the problem. A finite element basis over the domain Ω is used to discretize the infinite dimensional problem (2), such that $\mathbf{f}$ is the discretization of the nonparametric term f, Ψ is the matrix of the evaluation of the finite element basis functions at the observed spatial locations, P the discretization of the penalization term in (2), and $Q := I - \Psi(\Psi^\top \Psi)^{-1}\Psi^\top$. Then the estimators for $\boldsymbol{\beta}$ and $\mathbf{f}$ are provided by [10] and take the form

$$\begin{aligned} \hat{\mathbf{f}} &= (\Psi^\top Q \Psi + P)^{-1} \Psi Q \mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (X^\top X)^{-1} X^\top (\mathbf{y} - \Psi \hat{\mathbf{f}}). \end{aligned}$$

To convey a better understanding of the utility of the model, we examine the dataset depicted in Fig. 1, which shows the Chlorophyll-a concentration across the Mediterranean Sea. These data originate from the NASA Modis-Aqua satellite and are accompanied by measurements of sea surface temperature, as well as concentrations of particulate organic and inorganic matter at corresponding locations. In this context, the Mediterranean Sea serves as the focal area of interest, denoted as Ω, and is discretized using the mesh outlined in Fig. 2. We assume that the observed chlorophyll-a concentration y can be explained as the sum of a nonparametric component f and the effect of three available covariates: particulate inorganic carbon concentration x_1, particulate organic carbon concentration x_2 and the sea surface temperature x_3, as expressed in the SR-PDE model (1).

3 Novel Inference Approaches on Nonparametric Term

As we stress in Sect. 1, all the Wald-like approaches proposed so far for inference on f suffer a variance misspecification induced by the regularization in (2). To

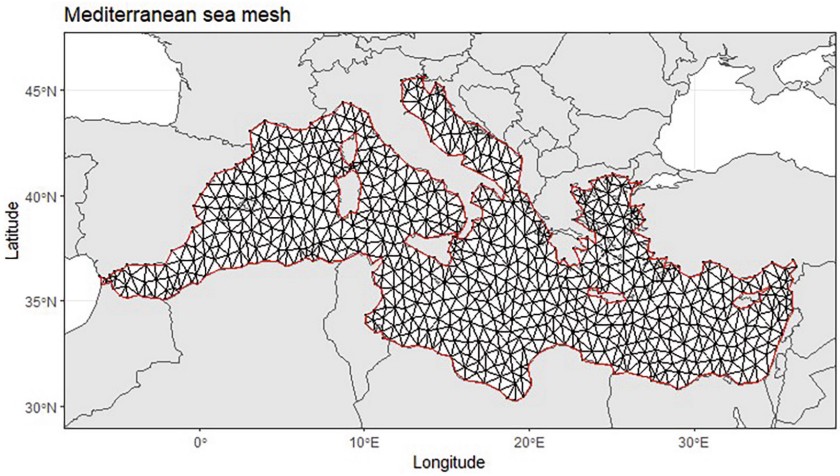


Fig. 2. Mediterranean sea mesh with 809 nodes.

address this problem, already encountered in literature, some authors suggest to undersmooth the model when performing inference (see, e.g., [12,14]). Following this idea, we propose to use bootstrap on the locations $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ to identify a proper scaling factor for the λ parameter in the functional (2), when estimating the variance of $\hat{\mathbf{f}}$. Specifically, we first obtain the estimate $\hat{\mathbf{f}}$ for $\mathbf{f}$ using the optimal λ selected via generalized cross validation. Then, we estimate the confidence intervals for $\mathbf{f}$ using a scaling factor on λ, checking whether the optimal estimate $\hat{\mathbf{f}}$ lies inside the confidence intervals. We then bootstrap the sample repeating the procedure each time, obtaining an estimated coverage of the confidence intervals. We hence evaluate a grid of user-defined scaling factors, selecting the one minimizing the distance between the nominal level and the estimated coverage of the confidence intervals. The resulting test and confidence intervals are less affected by the penalization, and therefore have better control of Type-I error. In turn, the bootstrap procedure has a relatively high computational cost.

For this reason we also propose a novel parametric method that relies on the asymptotic properties of the score equations of the model. The advantage of this line of proceeding, in analogy to the nonparametric tests in [3], is that the score equations are naturally independent from the effect of the penalization. Moreover, the usage of the asymptotic distributions of the score statistics contains the cost of the calculation of the confidence intervals, making them feasible also in the large sample scenarios. The resulting score statistic test has good statistical properties, while being computationally cheap.

The proposed tests can also be extended to space-time regression models with differential regularization, as those considered in [1,2].

4 Application to Chlorophyll-a Concentration in the Mediterranean Sea

The concentration of Chlorophyll-a is significant to understand the dynamics of the ocean carbon cycle and to assess the overall health of marine environments. Elevated levels of chlorophyll often indicate compromised water quality stemming from pollution and stagnation. The proliferation of microorganisms responsible for chlorophyll production is influenced by various environmental factors, including solar exposure, wind patterns, geographical features, temperature, salinity, and pollutants. While many of these factors can be directly measured at specific locations, and thus incorporated into models as spatially varying covariates, other environmental variables may not be measured locally. In such cases, these unmeasured factors can be collectively considered through a nonlinear spatial term.

In our specific application, we aim to model the chlorophyll-a concentration [15], as observed by the NASA satellite MODIS-AQUA, which provides data on a regular grid with a resolution of 9×9 kilometers. Additionally, measurements of inorganic [16] and organic [17] carbon concentration, along with sea surface temperature [18], are available at the same resolution and are considered as spatially varying covariates within an SR-PDE model, as specified in Sect. 2.

Our objective within this framework is to evaluate the significance of both the parametric and nonparametric components outlined in equation (1). The significance of the covariates can be assessed by means of the inference methods introduced in [7]. We then test the significance of the nonparametric term f by means of the proposed approaches, in addition to the Wald (parametric), the Sign-Flip and Eigen-Sign-Flip (nonparametric) tests described in [3]. At first, we test the significance of f on the whole domain Ω. All the tests agree on the significance of the spatial field, suggesting that the variability of the outcome cannot be explained solely through the available covariates, included in the parametric term. In addition, we also test the nonparametric term on some restricted spatial areas, for instance in the Thyrrenian sea. In this case we observe that the smooth field is indistinguishable from a constant field, as evidenced by all the methods employed. This is coherent with the fact that the unmeasured factors represented by f, such as sun exposure, winds, salinity, etc., vary substantially only on a larger geographical scale. Therefore if we restrict the analysis to a local scale, their effect is homogeneous, and gets overcome by the local factors represented in the considered covariates.

5 Conclusion

We propose two parametric tests for the smooth field f in SR-PDE models. The two tests have good control of the Type-I error, overcoming the existing parametric alternatives in the context. Moreover, the score asymptotic test is computationally cheap, enabling computation of confidence intervals also in the context of large data problems.

Acknowledgements. 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

1. Arnone, E., Azzimonti, L., Nobile, F., Sangalli, L.M.: Modeling spatially dependent functional data via regression with differential regularization. J. Multivar. Anal. **170**, 275–295 (2019)
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. Stoch. Env. Res. Risk Assess. **31**(1), 23–38 (2017)
3. Cavazzutti, M., Arnone, E., Ferraccioli, F., Galimberti, C., Finos, L., Sangalli, L. M.: Sign-flip tests for the nonparametric component in spatial regression with PDE regularization. In: Proceedings of the GRASPA 2023 Conference. Palermo University Press (2023)
4. Chen, H., Wang, Y., Li, R., Katherine, S.: A note on a nonparametric regression test through penalized splines. Stat. Sin. **24**, 1143–1160 (2014)
5. Crainiceanu, C., Ruppert, D., Claeskens, G., Wand, M.P.: Exact likelihood ratio tests for penalized splines. Biometrika **92**(1), 91–103 (2005)
6. Ferraccioli, F., Sangalli, L.M., Finos, L.: Nonparametric tests for semiparametric regression models. TEST **32**, 1106–1130 (2023)
7. Ferraccioli, F., Sangalli, L.M., Finos, L.: Some first inferential tools for spatial regression with differential regularization. J. Multivar. Anal. **189**, 104866 (2022)
8. Hemerik, J., Goeman, J., Finos, L.: Robust testing in generalized linear models by sign-flipping score contributions. J. Roy. Stat. Soc. B **82**, 841–864 (2020)
9. Liu, A., Wang, Y.: Hypothesis testing in smoothing spline models. J. Stat. Comput. Simul. **74**(8), 581–597 (2004)
10. Sangalli, L.M.: Spatial regression with partial differential equation regularisation. Int. Stat. Rev. **89**, 505–531 (2021)
11. Sangalli, L.M., Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. J. Roy. Stat. Soc. B **75**, 681–703 (2013)
12. Speckman, P.: Kernel smoothing in partial linear models. J. Roy. Stat. Soc. B **50**(3), 413–436 (1988)
13. Wood, S.N.: On p-values for smooth components of an extended generalized additive model. Biometrika **100**(1), 221–228 (2013)
14. Wood, S.N.: Generalized Additive Models. Chapman and Hall/CRC (2017)
15. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group: Moderate-resolution imaging spectroradiometer (MODIS) aqua chlorophyll data (2022). https://doi.org/10.5067/AQUA/MODIS/L3M/CHL/2022
16. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group: Moderate-resolution imaging spectroradiometer (MODIS) aqua particulate inorganic carbon data (2022)

17. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group: Moderate-resolution imaging spectroradiometer (MODIS) aqua particulate organic carbon data (2022)
18. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group: Moderate-resolution imaging spectroradiometer (MODIS) aqua sea surface temperature data (2022)

Bispectral Analysis of Markov Switching Bilinear Models

Maddalena Cavicchioli[1(✉)], Ahmed Ghezal[2], and Imane Zemmouri[3]

[1] Department of Economics Marco Biagi,University of Modena and Reggio Emilia, Viale Berengario 51, 41121 Modena, Italy
maddalena.cavicchioli@unimore.it
[2] Department of Mathematics, University Center of Mila, Mila, Algeria
[3] Department of Mathematics, University of Annaba, Annaba, Algeria
https://morespace.economia.unimore.it/maddalenacavicchioli/

Abstract. We derive matrix expressions in closed form for the spectral and bispectral densities of Markov switching bilinear models. Under suitable assumptions, we prove that the sample estimators of the spectral and bispectral density matrices are consistent and asymptotically normally distributed. Simulations and empirical applications confirm the validity of the asymptotic properties and the suitability of the proposed methods for the analysis of time series in the frequency domain.

Keywords: Markov switching models · bilinear processes · (bi)spectral density matrix · asymptotic theory

1 Introduction

Since the seminal works by Hamilton (1989, 1990) Markov switching (MS) time series models have received a growing interest in economics and statistics and continue to gain more popularity especially to adequately describe economic and financial data. An interesting class of MS time series models, called *Markov switching bilinear* (MS BL) models, has been studied by Bibi and Ghezal (2015, 2018) in which the process follows locally from a bilinear representation in the sense of Subba Rao *et al.* (1981, 1983, 1984).

In this paper we consider univariate MS BL models and we concentrate on some problems that were not completely solved in the literature. Specifically, we contribute to the literature in threefolds. First, we derive general and explicit matrix formulas in closed form for the autocovariance function of MS BL models. To do this, we use an indirect construction based on a suitable Markovian representation of the initial process. Second, based on the vector representation of the MS BL model, we compute the spectral and bispectral density matrices of the considered process, and show the importance of the (bi)spectral analysis in distinguishing the linear and bilinear models. Sample estimates of the spectral and bispectral densities are then proposed, and their consistency and asymptotic normality are established (under suitable assumptions). The main advantage of our

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 373–378, 2025.
https://doi.org/10.1007/978-3-031-64431-3_62

approach is related to the simplicity of the mathematical expressions that eases the computational effort. Third, we propose numerical and real-world examples to illustrate the obtained results, and to describe how to use the practical formulas in (bi)spectral analysis.

2 Markov Switching Bilinear Models

An $\mathbb{R}$-valued univariate process $(y_t)_{t\in\mathbb{Z}}$, defined on some probability space $(\Omega, \mathcal{F}, Pr)$, satisfies a *Markov switching subdiagonal bilinear model*, in short MS(M) SBL(p, q, p, Q), with $Q \leq p$, if it a solution of the following M-state stochastic difference equation

$$y_t \;=\; \sum_{i=1}^{p} a_i(s_t)\, y_{t-i} \;+\; \sum_{j=0}^{q} b_j(s_t)\, \eta_{t-j} \;+\; \sum_{j=1}^{Q}\sum_{i=j}^{p} c_{ij}(s_t)\, y_{t-i}\, \eta_{t-j} \qquad (1)$$

where (η_t) is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and unit variance, i.e., $\eta_t \sim \text{IID}(0,1)$, (s_t) is an irreducible and aperiodic Markov chain with a finite state space $\Xi = \{1, \ldots, M\}$, stationary transition probabilities $p_{ij} = Pr(s_t = j | s_{t-1} = i)$, collected in the $M \times M$ matrix $\mathbb{P}$, and unconditional (or steady state) probabilities $\pi_i = Pr(s_t = i)$, for $i \in \Xi$. The model coefficients $a_i(s_t)$, $b_j(s_t)$ and $c_{ij}(s_t)$ are functions of the unobserved Markov chain (s_t). Here $b_0(s_t) \neq 0$. In addition, assume that the chain (s_t) is stationary, and that (η_t) is independent of (s_t).

We consider a Markovian representation of (1) (set $n = p + q$ and $N = n\,(Q+1)$) which is given by

$$\mathbf{z}_t \;=\; \boldsymbol{\Phi}_t\, \mathbf{z}_{t-1} \;+\; \boldsymbol{\omega}_t \qquad (2)$$

where:

$$\mathbf{z}_t \;=\; (\mathbf{x}_t^{'}, \mathbf{x}_t^{'}\,\eta_t, \mathbf{x}_{t-1}^{'}\,\eta_{t-1}, \ldots, \mathbf{x}_{t-Q+1}^{'}\,\eta_{t-Q+1})^{'} \in \mathbb{R}^N$$

$$\mathbf{x}_t \;=\; (y_t, y_{t-1}, \ldots, y_{t-p+1}, \eta_t, \eta_{t-1}, \ldots, \eta_{t-q+1})^{'} \in \mathbb{R}^n$$

$$\boldsymbol{\omega}_t \;=\; \boldsymbol{\omega}(s_t, \eta_t) \;=\; (\mathbf{b}_t^{'}\,\eta_t, \mathbf{b}_t^{'}\,\eta_t^2, \mathbf{0}_{1\times n}^{'}, \ldots, \mathbf{0}_{1\times n}^{'})^{'} \in \mathbb{R}^N$$

$$\mathbf{b}_t \;=\; \mathbf{b}(s_t) \;=\; (b_0(s_t), 0, \ldots, 0, 1, 0, \ldots, 0)^{'} \in \mathbb{R}^n$$

and $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}(s_t, \eta_t)$ is the $N \times N$ matrix naturally deduced using the model coefficients in (1) in order to satisfy the stochastic matrix equation in (2).

3 Autocovariance Structure

Once second-order stationarity is ensured, it can be useful to compute the autocovariance function of $(\mathbf{z}_t)$ and (y_t). For computational convenience, we consider the centred version of the vector process $(\mathbf{z}_t)$, that is,

$$\mathbf{z}_t^* \;=\; \mathbf{\Phi}_t\, \mathbf{z}_{t-1}^* \;+\; \boldsymbol{\omega}_t^* \tag{3}$$

where $\mathbf{z}_t^* \;=\; \mathbf{z}_t \;-\; \boldsymbol{\mu}_{\mathbf{z}}$ and $\boldsymbol{\mu}_{\mathbf{z}} = E[\mathbf{z}_t]$. Then $\boldsymbol{\omega}_t^*$ is the centred residual vector such that, given $s_t = i \in \Xi$, $\boldsymbol{\omega}_t^*(i, \eta_t)$ is orthogonal to $\mathbf{z}_\tau^*$ for all $\tau < t$. By using (2) and (3), we get

$$\boldsymbol{\omega}_t^* \;=\; \boldsymbol{\omega}_t \;+\; (\mathbf{\Phi}_t \;-\; \mathbf{I}_N)\, \boldsymbol{\mu}_{\mathbf{z}}.$$

Then we have

Theorem 1. *Let us consider the* MS(M) SBL(p, q, p, Q) *model in* (1) *with Markovian representation* (2). *The autocovariance functions of* $(\mathbf{z}_t^*)$ *in* (3) *and its first component* (y_t^*) *are given by*

$$\begin{aligned}
\operatorname{vec} \mathbf{\Gamma}_{\mathbf{z}^*}(h) \;&=\; (\mathbf{e}^{'} \;\otimes\; \mathbf{I}_{N^2})\, \mathbf{\Pi}_{E[\mathbf{z}_t^* \otimes \mathbf{z}_{t-h}^* | s_t = \cdot]} \\
\gamma_{y^*}(h) \;&=\; \left(\mathbf{e}^{'} \;\otimes\; \boldsymbol{\delta}^{'} \otimes \boldsymbol{\delta}^{'}\right)\, \mathbf{\Pi}_{E[\mathbf{z}_t^* \otimes \mathbf{z}_{t-h}^* | s_t = \cdot]}
\end{aligned}$$

where

$$\mathbf{\Pi}_{E[\mathbf{z}_t^* \otimes \mathbf{z}_{t-h}^* | s_t = \cdot]} \;=\; \left(\mathbb{P}_{\mathbf{\Phi}}^h \;\otimes\; \mathbf{I}_N\right)\, \left(\mathbf{I}_{MN^2} \;-\; \mathbb{P}_{\mathbf{\Phi}^{(2)}}\right)^{-1}\, \mathbf{\Pi}_{\boldsymbol{\omega}^{*(2)}}$$

for every $h \geq 0$ *(set* $\mathbf{A}^h = \mathbf{I}_m$ *if* $h = 0$ *for any square* $m \times m$ *matrix* $\mathbf{A}$*). Furthermore,* $\mathbf{\Gamma}_{\mathbf{z}^*}(h) = \mathbf{\Gamma}_{\mathbf{z}^*}^{'}(-h)$ *and* $\gamma_{y^*}(h) = \gamma_{y^*}(-h)$ *for every* $h < 0$. *Here we set* $\mathbf{e} = (1, \cdots, 1)^{'} \in \mathbb{R}^M$ *and* $\boldsymbol{\delta} = (1, 0, \cdots, 0)^{'} \in \mathbb{R}^N$.

In the above statement $\mathbb{P}_{\mathbf{\Phi}}$ is the square matrix whose (i, j)-element is $p_{ji}\mathbf{\Phi}(i)$, with $\mathbf{\Phi}(i) = E_\eta[\mathbf{\Phi}_t | s_t = i]$. Moreover, $\mathbb{P}_{\mathbf{\Phi}^{(2)}}$ is defined in the same manner as $\mathbb{P}_{\mathbf{\Phi}}$ substituting $\mathbf{\Phi}$ by the Kronecker product $\mathbf{\Phi} \otimes \mathbf{\Phi}$. Finally, $\mathbf{\Pi}_{\boldsymbol{\omega}^{*(2)}}$ is the matrix whose i-row is given by $\pi_i \boldsymbol{\omega}^{*(2)}(i)$, with $\boldsymbol{\omega}^{*(2)}(i) = E_\eta[\boldsymbol{\omega}_t^* \otimes \boldsymbol{\omega}_t^* | s_t = i]$.

4 Spectral and Bispectral Density Matrices

The spectral density $\mathbf{F}_{\mathbf{z}^*}$ of $(\mathbf{z}_t^*)$ can be obtained from $\mathbf{\Gamma}_{\mathbf{z}^*}(h)$ by taking its Fourier transform, that is,

$$\operatorname{vec} \mathbf{F}_{\mathbf{z}^*}(\omega) \;=\; \sum_{h=-\infty}^{\infty} \operatorname{vec} \mathbf{\Gamma}_{\mathbf{z}^*}(h)\, e^{-i\omega h} \tag{4}$$

where the frequency ω belongs to $[-\pi, \pi]$. The diagonal terms of the spectral density matrix $\mathbf{F}_{\mathbf{z}^*}(\omega)$ are computed by considering the summation (also denoted by the same symbol)

$$\operatorname{vec} \mathbf{F}_{\mathbf{z}^*}(\omega) \;=\; \sum_{h=-\infty}^{\infty} \operatorname{vec} \mathbf{\Gamma}_{\mathbf{z}^*}(|h|)\, e^{-i\omega h}$$

which can be expressed as

$$\operatorname{vec}\mathbf{F}_{\mathbf{z}^*}(\omega) \;=\; \left(\mathbf{e}^{'}\otimes\mathbf{I}_{N^2}\right)\left[\sum_{h=-\infty}^{\infty}\left(\mathbb{P}_{\boldsymbol{\Phi}}^{|h|}\;\otimes\;\mathbf{I}_N\right)e^{-i\omega h}\right] \times\;\left(\mathbf{I}_{MN^2}\;-\;\mathbb{P}_{\boldsymbol{\Phi}^{(2)}}\right)^{-1}\boldsymbol{\Pi}_{\boldsymbol{\omega}^{*(2)}} \tag{5}$$

from Theorem 1 above. This is a neat matrix expression in closed form that eases the computational effort to derive the spectral density. The above series converges as $\rho(\mathbb{P}_{\boldsymbol{\Phi}^{(2)}})<1$, which also implies that $\rho(\mathbb{P}_{\boldsymbol{\Phi}})<1$, where $\rho(\cdot)$ is the spectral radius of the considered square matrix.

Theorem 2. *The spectral density matrix of the centred process* $(\mathbf{z}_t^*)$ *in* (3) *is given by*

$$\operatorname{vec}\mathbf{F}_{\mathbf{z}^*}(\omega) \;=\; \left(\mathbf{e}^{'}\otimes\mathbf{I}_{N^2}\right)\left[-\mathbf{I}_{MN^2}\;+\;2\,\operatorname{Re}\mathbf{X}(\omega)\right]\mathbf{W}$$

where

$$\mathbf{X}(\omega) \;=\; \left(\mathbf{I}_{MN}\;-\;\mathbb{P}_{\boldsymbol{\Phi}}\,e^{-i\omega}\right)^{-1}\otimes\mathbf{I}_N$$
$$\mathbf{W} \;=\; \left(\mathbf{I}_{MN^2}\;-\;\mathbb{P}_{\boldsymbol{\Phi}^{(2)}}\right)^{-1}\boldsymbol{\Pi}_{\boldsymbol{\omega}^{*(2)}}.$$

For the centred process (y_t^*)*, defined as the first component of* $(\mathbf{z}_t^*)$*, we get*

$$\gamma_{y^*}(\omega) \;=\; \left(\mathbf{e}^{'}\otimes(\boldsymbol{\delta}^{'})^{\otimes 2}\right)\left[-\mathbf{I}_{MN^2}\;+\;2\,\operatorname{Re}\mathbf{X}(\omega)\right]\mathbf{W}$$

where the frequency ω *belongs to* $[-\pi,\pi]$*.*

Furthermore, using the third ordinary cumulant, define

$$\operatorname{vec}\mathbf{C}_{\mathbf{z}^*}(h,k) \;=\; E[\mathbf{z}_t^*\;\otimes\;\mathbf{z}_{t-h}^*\otimes\mathbf{z}_{t-k}^*].$$

Taking the Fourier transform yields the bispectral density function of $(\mathbf{z}_t^*)$, that is,

$$\operatorname{vec}\mathbf{F}_{\mathbf{z}^*}(\omega_1,\omega_2) \;=\; \sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\operatorname{vec}\mathbf{C}_{\mathbf{z}^*}(h,k)\,e^{-ih\omega_1-ik\omega_2}$$

for $\omega_1,\omega_2\in[-\pi,\pi]$. Reasoning as above, and to simplify computation we consider the following summation (also denoted by the same symbol):

$$\operatorname{vec}\mathbf{F}_{\mathbf{z}^*}(\omega_1,\omega_2) \;=\; \sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\operatorname{vec}\mathbf{C}_{\mathbf{z}^*}(|h|,|k|)\,e^{-ih\omega_1-ik\omega_2}.$$

We now derive a matrix formula in closed form for the bispectral density function defined above. For every $h,k>0$ with $h<k$ (the case $h\geq k$ can be derived in a similar manner), we get

$$\operatorname{vec}\mathbf{C}_{\mathbf{z}^*}(h,k) \;=\; \left(\mathbf{e}^{'}\otimes\mathbf{I}_{N^3}\right)\boldsymbol{\Pi}_{E[\mathbf{z}_t^*\otimes\mathbf{z}_{t-h}^*\otimes\mathbf{z}_{t-k}^*|s_t=\cdot]}$$

where

$$\mathbf{\Pi}_{E[\mathbf{z}_t^* \otimes \mathbf{z}_{t-h}^* \otimes \mathbf{z}_{t-k}^* | s_t = \cdot]} = \left(\mathbb{P}_{\mathbf{\Phi}}^h \otimes \mathbf{I}_{N^2}\right) \left(\mathbb{P}_{\mathbf{\Phi}^{(2)}}^{k-h} \otimes \mathbf{I}_N\right) \mathbf{\Pi}_{E[\mathbf{z}_t^{*\otimes 3} | s_t = \cdot]}.$$

Sample estimates of the spectral and bispectral densities are then proposed, and their consistency and asymptotic normality are established (under suitable assumptions).

5 An Empirical Application

We consider the IBM stock closing price data as in Aknouche and Rabehi (2010). Such data are subject to multimodality and nonlinear features. In Fig. 1 we report the bispectral density plot of IBM stock prices. From this graph we recognize that bimodality is actually present at lower frequencies, indicating recurrent long cycles and suggesting a switching recursive model at the frequencies of the business cycle.

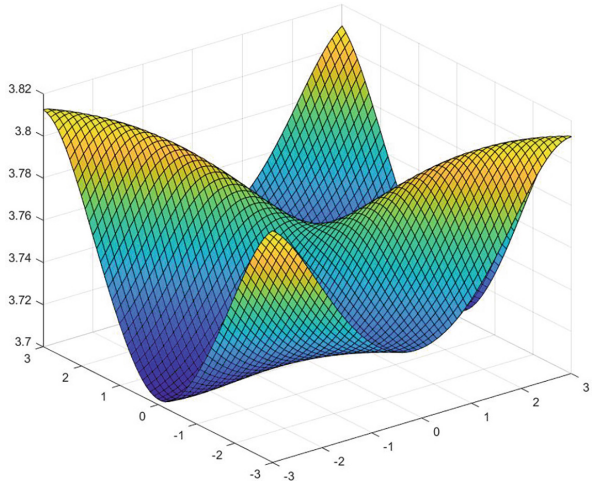


Fig. 1. Bispectral density function of the IBM stock prices mixture bilinear model in Aknouche and Rabehi (2010).

References

Aknouche, A., Rabehi, N.: On an independent and identically distributed mixture bilinear time series model. J. Time Ser. Anal. **31**(2), 113–131 (2010)

Bibi, A., Ghezal, A.: On the Markov-switching bilinear processes: stationarity, higher-order moments and β-mixing. Stochast. Int. J. Probab. Stochast. Processes **87**(6), 919–945 (2015)

Bibi, A., Ghezal, A.: Markov-switching bilinear GARCH models: structure and estimation. Commun. Stat. Theory Methods **47**(2), 307–323 (2018)

Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica **57**, 357–384 (1989)

Hamilton, J.D.: Analysis of time series subject to changes in regime. J. Econometrics **45**, 39–70 (1990)

Rao, T.S.: On the theory of bilinear time series model. J. Roy. Stat. Soc. Ser. B Stat. Methodol. **43**(2), 244–255 (1981)

Subba Rao, T.: The bispectral analysis of nonlinear stationary time series with reference to bilinear time series models. In: Brillinger, D.R., Krishnaiah, P.R. (eds.) Handbook of Statistics, vol. 3, pp. 293–319. North-Holland, Amsterdam, New York, Oxford (1983)

Rao, T.S., Gabr, M.M.: An introduction to bispectral analysis and bilinear time series models. Lecture Note in Statistics, vol. 24, Springer, Verlag-New York-Berlin-Heidelberg-Tokyo (1984)



Topological Data Analysis Applied to Radiomics (topiomics) Data in Recurrent Prostate Cancer

Lara Cavinato[1(✉)], Lorenzo Ferrara[1], Matteo Pegoraro[2], Paola Anna Erba[3], and Francesca Ieva[1,4]

[1] MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy
lara.cavinato@polimi.it
[2] Department of Mathematical Science, Aalborg University, Aalborg, Denmark
[3] Department of Medicine and Surgery, University of Milan Bicocca, Milan, Italy
[4] Health Data Science centre, Human Technopole, Milan, Italy

Abstract. Non-invasive imaging techniques, particularly Positron Emission Tomography (PET), have revolutionized medical diagnostics. Radiomics, an emerging field, aims to extract quantitative data from tomographic images to reveal subtle tissue alterations. Despite challenges in standardization, radiomics holds promise in linking imaging features to clinical outcomes, crucial in metastatic cancer treatment. This study explores Topological Data Analysis (TDA) to dissect intra-tumor heterogeneity in metastatic prostate cancer. Patient representation methods, including hierarchical dendrograms and persistence diagrams, are compared. Significant differences in clinical variables and treatment responses are observed across patient clusters. This research contributes to advancing precision oncology in metastatic cancer subtyping.

Keywords: topological data analysis · radiomics · prostate cancer

1 Introduction

Non-invasive imaging methods like Positron Emission Tomography (PET) have long been essential tools for medical professionals, aiding in the analysis, diagnosis, and treatment of various diseases. Radiomics, a burgeoning field, enhances traditional diagnostic techniques by extracting quantitative data from tomographic images, revealing subtleties in tissue alterations and pathological processes not visible to the naked eye. Despite technical challenges stemming from inconsistent imaging protocols and quality assurance measures, radiomics holds promise in establishing links between imaging features and clinical outcomes. Recent studies have increasingly focused on radiomic variables, recognizing tumor heterogeneity as crucial for treatment response and prognosis [1]. Addressing challenges posed by metastatic cancers, which exhibit diverse responses to therapy across multiple lesions, researchers advocate for a comprehensive approach encompassing both primary and metastatic lesions to capture the full

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 379–385, 2025.
https://doi.org/10.1007/978-3-031-64431-3_63

tumor heterogeneity spectrum [2,3]. Building on these advancements, our study seeks to delve deeper into the complexities of intra-tumor heterogeneity using advanced analytical techniques in the range of statistical methods. Specifically, we aim to leverage Topological Data Analysis (TDA) to characterize the nuanced structural patterns within tumors, transcending traditional clustering methodologies to provide a more holistic understanding of tumor biology [4]. Building upon [5], we propose to extend the application of TDA to analyze the whole sets of lesions rather than individual ones, thereby capturing the collective heterogeneity within a tumor. TDA offers a unique perspective by considering the topological nature of data, enabling the identification of shape patterns and structural features that may hold clinical significance. Examples of TDA applications involve the use of tree-like structures, such as phylogenetic trees and hierarchical clustering dendrograms, or the use of other topological summaries such as persistence diagrams, persistence silhouettes, and persistence landscapes. In this work, we aim at comparing two different patient representation strategies, hierarchical dendrograms and persistence diagrams, in order to stratify patients affected by Recurrent, i.e., metastatic, Prostate Cancer (PCa). By embracing a multidimensional perspective, we can move closer to realizing the vision of precision oncology, where treatment strategies are tailored to the individual characteristics of each patient's tumor.

2 Methods

In this section we provide details regarding the dataset employed for the analyses and the data preoprocessing (Sect. 2.1), the patient representation strategies (Sect. 2.2), and the patient stratification pipelines (Sect. 2.3).

2.1 Dataset

Data from 47 Biomchemical Recurrent Prostate Cancer (BRPCa) patients who underwent PET scans using [68Ga]PSMA-11 tracer were analyzed. Patient information included age, lesion site, Gleason Score, PSA levels, number of lesions, and therapy history. PSA levels were recorded before and after Primary Treatment (TP), during hormonal therapy PET scans, and at Biochemical Recurrence (BR). While PSA is a common marker for prostate conditions, its variations can be influenced by factors beyond cancer. Evaluating disease progression based on PSA level variations relative to initial levels is important, as absolute differences may be misleading due to inherent baseline variability [6]. PET/CT image analysis yielded 129 radiomic features (specifically from the radiomic groups of GLSZM, GLCM, GLRLM, and NGTDM) per lesion, but a high rate of missing values, particularly in certain radiomic groups, was noted, possibly due to technical challenges affecting data reliability.

2.2 Patient Representation

Radiomics features were considered for patient representation and, specifically, were reduced with Principal Component Analysis (PCA) before usage. Being

q the number of principal components and n_i the number of lesion belonging to patient i, we define patient i as $C_i = x_1^q, ..., x_{n_i}^q$, where x_j^q with $j = 1, ..., n_i$ are the lesions' representation in the radiomics space $\mathbb{R}^q$. The set is said point cloud. Two different methods from Topological Data Analysis were employed to summarize the patients' point clouds: hierarchical dendrograms and persistence diagrams.

Hierarchical Dendrograms. Hierarchical clustering dendrograms have been proposed as a summary as they unveil the intrinsic relationship among points of a point cloud. A dendrogram is obtained in such a way that lesions are linked in terms of similarity relationship in their imaging characteristics. It quantifies to which extent lesions, i.e., their radiomic vectors, are similar within patients and how they get agglomerated hierarchically, one to each other. Operationally, we computed the pairwise Euclidean distances between peer lesions to generate a distance matrix, used to build a dendrogram with average linkage. The structure of the dendrogram visually represents the merging patterns within the data, reflecting the distribution of the intra-tumor variability [5].

Persistence Diagrams. A distinct topological method was employed, treating each patient's lesion cloud point as a Persistence Diagram (PD). In contrast to trees, PDs are part of the broader framework of Persistent Homology, a TDA technique for assessing topological data features across various spatial resolutions. This involves creating a series of topological spaces based on increasing disk radii, known as a filtration, to analyze shape characteristics at different scales. From this filtration, homology groups are extracted to describe the evolution of topological features. Dendrograms visually represent path-connected components at a fixed height, akin to clusters formed by merged points. In comparison, Persistence Diagrams condense filtration information into points $x_1, ..., x_N \in \mathbb{R}^2$ representing the appearance and disappearance of topological features, such as connected components, holes, or voids. Our focus on lesion grouping patterns led to the use of persistence diagrams, particularly highlighting connected components, where features start at a radius ρ of 0 and merge as the radius expands, symbolizing the merging of disks when their centers are within a certain distance [7]. In practical terms, the merging of two disks occurs when their centers are positioned in a way that $d(O_1, O_2) < 2\rho$.

2.3 Patient Stratification and Cluster Interpretation

To perform clustering we need a proper distance, which capture the different extent of tumor heterogeneity and allows the pair-wise comparison between patient-representation object, i.e., hierarchical trees and persistence diagrams.

Clustering of Hierarchical Trees. After obtaining patients' trees, they were compared using an Edit Distance (d_E) metric. To compute the d_E measure

between two trees, changes are made to the structure of the first tree to align it topologically with the other. These modifications include shrinking, inserting, or deleting edges, with their impact on distance computation contingent upon the length of the affected edge. This method establishes a distance metric between trees by identifying the shortest path, requiring minimal alterations, to convert the first tree into the other. To enhance the clustering procedure's robustness, a regularization approach was implemented through a Pruning Edit distance (PTED) [5]. Utilizing this method, leaf pairs with reciprocal distances below a selected threshold ϵ are deemed sufficiently homogeneous to be merged into a single lesion. Thus, this approach contrasts with mere leaf counting by considering distinct phenotypes identified through pruning, effectively mitigating the influence of minor differences in lesion phenotypes on the overall structure while preserving significant heterogeneity. However, due to PTED's sensitivity to the ϵ parameter, an averaging approach was employed using multiple ϵ values, each weighted according to a chosen distribution μ. This technique redefines PTED, facilitating the characterization of both pruning intensity and focus, thereby enhancing analytical flexibility and robustness. In our study, a Beta parametric model was employed to select the μ distribution, refined by comparing it with the merging heights of patients' trees to reveal the rate of small leaf merging into larger clusters and highlight variability and heterogeneity patterns within the dataset.

Clustering of Persistence Diagrams. Once we acquired the persistence diagram for each patient, we established the distance between two diagrams through a matching, where each point in S is paired with a point in T or on the diagonal using a bijection $\phi : S \to T \cup (x, x) : x \in \mathbb{R}$. The distance is determined by evaluating the difference between each point and its corresponding match, calculated using a cost function. In our analysis, we utilized the q-Wasserstein distance and the Bottleneck distance, named for their associations with partial optimal transport. For $q \geq 1$, the Wasserstein metric is conventionally defined, but for $q < 1$, an extension is employed by introducing a new function $\tilde{d}_q(x, y) = (d_q(x, y))^q$, which maintains coherence with the interpretation of the classical q-Wasserstein metric. The Bottleneck distance, representing the supremum of distances and considering only the largest distance, can be viewed as the limit of the Wasserstein distance as q approaches infinity. Subsequently, the obtained distance matrices underwent the clustering algorithm, simultaneously utilizing multiple distance matrices derived from both the Bottleneck distance and the Wasserstein distance, incorporating various values of q within the range $(0, 3]$. The resulting clusters were then analyzed and characterized [8].

2.4 Cluster Analysis

Following the computation of distances between patients, Hierarchical Clustering was employed, with the number of clusters set to 3 to align with the dendrogram representation. Clinical features were utilized to characterize the phenotype and

treatment plan of the identified clusters. Mann-Whitney non-parametric tests were utilized for comparing distributions of continuous variables, while Chi-squared tests were employed for assessing independence of categorical variables. Pairwise one-sided comparisons between groups were conducted instead of multivariate analysis to provide a group-wise characterization, with a significance level of $\alpha = 0.1$ selected to address the impact of small sample size.

3 Results

3.1 Dendrograms-Based Pipeline

The Hierarchical clustering with complete linkage resulted in imbalanced cluster sizes, comprising 27, 7, and 1 elements, respectively. The last cluster was considered an outlier and subsequently discarded, while the first two were subjected to further analyses. Specifically, we examined the clusters to understand how clinical variables were influenced by and interacted with the clustering procedure. We assessed disparities between the derived groups regarding response to treatment (refractory vs. relapsing), initial PSA values and their changes over time, number of lesions, age, and Gleason score. Additionally, we evaluated the correlation between the clustering patterns and other clinical information, such as the implementation of specific therapies. In the dataset, significant differences were found for age ($p = 0.082$), number of lesions ($p < 0.001$), PSA levels variation during the PET ($p = 0.006$), and Gleason Scores, which describe the aggressiveness of the metastatic disease. Particularly, when testing the GS categorical variable, an Independence test was employed ($p = 0.06$). However, due to the limited number of observations per group, a numerical approach was also utilized, involving consideration of the numerical result C for each Score of the form of A + B = C and conducting tests based on these numerical values. This test was consistent with the previous one ($p = 0.001$); however, it's important to note that the labels A + B and B + A are not clinically equivalent. Moreover, the clusters showed significant differences in the number of lesions ($p < 0.001$), initial PSA levels ($p = 0.083$), and PSA levels variation after the TP ($p = 0.091$) and during the PET ($p = 0.006$). The clustering procedure also revealed differences in treatment therapies, with LHRHa and Abiraterone Acetate (AA) emerging as the main differentiating factors ($p = 0.06$ and $p = 0.04$, respectively).

3.2 Persistence Diagrams-Based Pipeline

Similar to the previous section, the Persistence homology approach also categorizes the patients into three clusters. Cluster 2 contained a limited number of observations, resulting in the classification of elements within this cluster as outliers. Consequently, only clusters 0 and 1 were subjected to analysis. From a clinical perspective, the characterization of the clusters aligned with the one performed in the PTED framework. Specifically, in both datasets, significant differences were observed in the variations of PSA levels (After treatment: p =

0.04; During PET: p = 0.002) and in the number of lesions (p = 0). Similarly, a difference in the value of the Gleason score is observed only in the dataset (p = 0.05). The intuitive parallelism between the results yielded by the PTED approach and the PD approach was emphasized by directly comparing the clustering labels. A high level of concordance was evident and could be quantified through the use of the F1 score, which was 0.94.

4 Conclusions

Interpreting radiomics features extracted from PET/CT exams remains a significant challenge, and ongoing research is dedicated to establishing a robust mathematical and statistical foundation to aid treatment decision-making for patients with multiple lesions. In this study, we addressed this challenge by exploring TDA methods to leverage the intrinsic information captured by radiomic features in multi-lesion patients. Our objective was to uncover significant correlations between radiomic signatures detected in the patients under study and the distinctive characteristics observed in the phenotype and evolution of their corresponding cancers, with particular emphasis on characterizing cancer heterogeneity.

Acknowledgements. We acknowledge all the personnel of Medicine Department of Azienda Ospedaliero-Universitaria Pisana for the assistance during the PET/CT scans, segmentation of lesions, extraction of radiomic features and retrieval of patients' personal information from EHR. We particularly thank dr. Paola Anna Erba and dr. Costanza Bachi for their support.

L. Cavinato is funded by the National Plan for NRRP Complementary Investments - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM).

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

References

1. Majumder, S., et al.: State of the art: radiomics and radiomics related artificial intelligence on the road to clinical translation. BJR Open **6**(1), tzad004 (2023)
2. Chang, E., et al.: Comparison of radiomic feature aggregation methods for patients with multiple tumors. Sci. Rep. **11**(1), 9758 (2021)
3. Cavinato, L., et al.: Radiomics-based inter-lesion relation network to describe [18F] FMCH PET/CT imaging phenotypes in prostate cancer. Cancers **15**(3), 823 (2023)
4. Chazal, F., Michel, B.: An introduction to topological data analysis: fundamental and practical aspects for data scientists. Front. Artif. Intell. **4**, 108 (2021)
5. Cavinato, L., et al.: Imaging-based representation and stratification of intra-tumor heterogeneity via tree-edit distance. Sci. Rep. **12**(1), 19607 (2022)
6. Simon, N.I., et al.: Best approaches and updates for prostate cancer biochemical recurrence. Am. Soc. Clin. Oncol. Educ. Book **42**, 352–359 (2022)

7. Divol, V., Lacombe, T.: Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport. J. Appl. Comput. Topology **5**, 1–53 (2021)
8. Givens, C.R., Shortt, R.M.: A class of Wasserstein metrics for probability distributions. Mich. Math. J. **31**(2), 231–240 (1984)



PixHomology: A New Algorithm for Computing Persistent Homology for Pixel Data

Riccardo Ceccaroni(✉) and Pierpaolo Brutti

Department of Statistical Sciences, Università di Roma "La Sapienza", P.le Aldo Moro 5, Rome 00185, Italy
riccardo.ceccaroni@uniroma1.it

Abstract. In this work, we present a highly efficient algorithm for zero-dimensional Persistent Homology (PH) computation on images. Our approach hinges on a novel method for filtering simplicial complexes derived from pixel connections. Through comparative analyses with established adjacency matrix-based methodologies, we show that our algorithm, *PixHomology*, drastically reduces memory usage while also improving computational speed. It is important to notice that the very small memory footprint readily allows to process high-resolution images across a wide range of application areas ranging from astronomy to histology.

Keywords: Persistent Homology · Image Analysis · Computational Efficiency · Simplicial Complexes

1 Introduction

Persistent Homology (PH) is a powerful mathematical tool used in topological data analysis (TDA) that has found wide application in the analysis of image arising from a variety of applied fields. At its core, PH, provides a way to extract the connectivity structure underlying the dataset of interest.

More specifically, in the context of image processing, PH allows us to extract information about the underlying topological features, such as connected components, and holes that can then be used to tackle different learning tasks such as object recognition, image segmentation, and pattern detection, where traditional methods may struggle to capture the full complexity of the data.

To date, a plethora of methodological advancements and software frameworks have emerged to facilitate the efficient computation of PH on images. Among these, *Ripser* [2] can be considered a reference pipeline mainly for its versatility and efficiency. *Ripser* in facts, stands out as the only software that employs simplicial complexes [12] for PH calculation on images. Other notable options include Dionysus [6], utilizing the Freudenthal triangulation [5] on a grid, and *DIPHA* [15], pioneering distributed PH computation using cubical complexes

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 386–392, 2025.
https://doi.org/10.1007/978-3-031-64431-3_64

[15]. More recently, alternative high-performance software solutions like CubicalRipser [8] and Cubicle [14] have emerged, focusing on PH computation on images via cubical complexes.

In this study, we focus on PH computation on images via simplicial complexes. In the following sections, we first introduce a novel algorithm designed for computing zero-dimensional PH, and then we conduct a comparative analysis with *Ripser*, evaluating both execution time and memory footprint.

2 PH Calculations on Images

Given a $2D$ grey-level image, the digital image function is defined by $\psi : \mathbb{Z}^2 \mapsto \mathbb{R}$. Since the co-domain of ψ is ordered, an ordered set of level lines on ψ can be obtained by taking values at $\{\ell_0, \ldots, \ell_i\}$ such that $\{\ell_0, \leq \ell_1 \leq \cdots \leq \ell_i\}$ with all level lines having real values in the range of ψ. The image can be represented by a collection of projections at values $H_f = \{\ell_0, \ldots, \ell_i\}$. Connected components at each level line ℓ_i, namely $CC(\ell_i)$.

By changing the height of the slicing plane from ℓ_0 to ℓ_i in the range of ψ, connected components gradually split and disappear according to the morphology of the elevation surface. The progressive changes of connected components show the inclusion relations on which the tree is constructed.

Such inclusion relations have a natural representation in the form of hierarchical structures. The hierarchical tree representing elements of CC and their inclusion relations is called an inclusion tree. In this representation, the root node is a unique connected component cc_0 that covers the entire image. The cc_0 matches a level line at the lowest value of ψ. Conversely, the leaves of the tree are local maxima of ψ.

The leaf nodes of a tree are likely to correspond to small disjointed areas in the image. Although small, they are local maxima, thus salient regions of ψ. Leaf nodes are considered representatives of the branch they belong to. A bifurcation is where two or more leaf nodes join. The root of the tree is the union of all leaf nodes. All nodes in the middle are progressively removed from the tree connecting each pixel with the pixels found among its 8 neighbors, until only leaf nodes, bifurcation nodes, and the tree root remain. Nodes maintain information about the height of the original level line.

Traversing the tree structure from the leaf node to the root produces various configurations of sets and subsets of leaf nodes. This process offers a way to discern the *birth* and *death* of the connected components. The *birth* value refers to the value of the local maximum found on each leaf. Whenever two leaves combine at a level l_i this is called the *death* value of the component with the lower *birth* value. Following this construction, the leaf containing the absolute maximum value will exhibit the *death* value at level l_0.

This approach defines the zero-dimensional PH. The *birth* and *death* values associated with each component can be visualized in a diagram known as the persistence diagram (PD). The difference between the *birth* and *death* values is termed the lifetime of the connected component.

The idea outlined here, although quite natural, might pose significant implementation problems. Specifically, the most direct approach to filter the ψ function involves generating an adjacency matrix. This matrix grows exponentially fast with the image size, presenting a significant challenge when dealing with large-scale images. Such images are increasingly common in applied fields like astronomy and histology, where dimensions can easily exceed $20,000 \times 20,000$ pixels. To mitigate excessive memory consumption and computational runtime, we now describe a novel algorithm: *PixHomology*. This algorithm aims to drastically cut both, memory usage and computation time.

3 PixHomology

Our algorithm comprises the five steps detailed below. The key idea to reduce memory usage involves bypassing the construction of the adjacency matrix. Instead, we directly identify the connected components and the pixels that connect them within the image. Subsequently, we generate the PD by exclusively analyzing the pixels we just identified.
Hence, given an input image I, the proposed algorithm consists of the following steps:

Step 1: ***Make pixels unique***
Since an image might have two neighboring pixels with the same value, we create an array N with a size equal to I initialized with random values from Gaussian distribution (mean 0, variance 1). In this way, if we take the pair of values from I and N with the same index, we are sure that it is unique.

Step 2 : ***Identification of the concave components***
This step is about the identification of the concave components within the image I. To do this we calculate, for each pixel, the index of the pixel with the highest value among the 8 neighboring pixels and store it in the M array. If there are two or more pixels in I with the highest value, we store in M the index that has the largest value in N.
Next, each element $x \in M$ is replaced by the value $M[x]$, which corresponds to the value found in position x of M. This process ends when x equals $M[x]$ for every element $x \in M$.

Step 3 : ***Edge points detection***
After dividing the image into different components, we identify the pixels in the border of the components. We insert the indices of these pixels into the array B.

Step 4 : ***Sorting***
We sort the indices contained in B. For each $b \in B$ we use the corresponding index in M to find the value in image I. The array B is sorted to have these values extracted from I in descending order.

Step 5 : ***Filtering and PD construction***
For all b in B we identify the two components in M linked by b. In particular, the component with the smallest local maximum value in I will be merged

with the other. This results in the inclusion of a *birth* point in the PD, i.e. the smallest local maximum, and a *death* point, i.e. the value of I corresponding to the index b. We repeat for all indices in B. If two components have already been merged, we continue with the next index b.

Algorithm 1 PixHomology Algorithm

Input: Image I
Output: PD
// Step 1: Make pixels unique
$N \sim$ Gaussian noise (mean 0, variance 1), N have the same size of I;
// Step 2: Identification of the concave components
$M = find_local_maxima(I, N)$;
while $x = M[x]\ \forall x$ **do**
    for $x \in M$ **do**
        $x = M[x]$
    end for
end while
// Step 3: Edge points detection
$B = find_union_pixels_index(M)$;
// Step 4: Sorting
$B = qsort(B, M, I)$;
// Step 5: Filtering and PD construction
Initialize empty PD
for b in B **do**
    Add *birth* point and *death* point of the component to PD
    Merge components in M, keeping component with highest local maximum
end for

4 Results

Since the only other software known to use simplicial complexes on images is *Ripser*, here it will serve as the benchmark for comparison with *PixHomology*. The comparison is performed on the following datasets:

- MNIST: 60.000 images with a size of 28×28 pixels [11];
- CIFAR10: 50.000 images with a size of 32×32 pixels [10];
- IMAGENET_A: 7.500 images with various sizes, cropped with size of 300×300 pixels [7];
- DIV2K: 800 images with various sizes, cropped with a size of 1.000×1.000 pixels [1];
- KATHER: 10 images with size of 5.000×5.000 pixels [9];
- MAST: 37 astronomical images from MAST repository with a size of 5.900×5.900 pixels.

The RGB images underwent conversion to grayscale before analysis. Additionally, while *PixHomology* supports the analysis of *float64* data, it was necessary to transform images with *float64* values to *float32*, aligning with the data format returned by *Ripser*.

Initially, we conducted a validation process to ensure consistency between the PD generated by *PixHomology* and *Ripser*. This verification involved calculating the bottleneck distance between the two diagrams, revealing consistent results across all examined images. Subsequently, we proceeded to assess the execution time, shown in Table 1, and memory utilization, shown in Table 2, of both algorithms. The source code of *PixHomology* to reproduce these experiments is available on https://github.com/riccardoc95/PixHomology.

Table 1. Execution time in seconds

Dataset	Method	min	mean	max
MNIST	*Ripser*	0.0024	0.0027	0.0196
	PixHomology	0.0003	0.0004	0.0025
CIFAR10	*Ripser*	0.0028	0.0031	0.0075
	PixHomology	0.0002	0.0004	0.0019
IMAGENET_A	*Ripser*	0.1307	0.1664	0.2269
	PixHomology	0.0185	0.0351	0.0503
DIV2K	*Ripser*	2.6221	3.1774	3.7344
	PixHomology	0.3974	1.4954	2.9368
KATHER	*Ripser*	148.4023	152.5039	156.1330
	PixHomology	8.0033	8.3768	8.5928
MAST	*Ripser*	134.2367	157.2324	159.6529
	PixHomology	12.7651	16.7098	17.7098

Table 2. Memory usage in Mb

Dataset	Method	min	mean	max
MNIST	*Ripser*	0.6162	0.6163	0.6337
	PixHomology	0.0044	0.0045	0.5009
CIFAR10	*Ripser*	0.8084	0.8093	0.8136
	PixHomology	0.0053	0.0055	0.4453
IMAGENET_A	*Ripser*	53.3070	53.5011	53.8053
	PixHomology	0.2634	0.2656	0.5627
DIV2K	*Ripser*	816.9848	819.4173	822.2476
	PixHomology	4.0014	4.0057	4.0061
KATHER	*Ripser*	20465.9248	20475.2238	20481.9552
	PixHomology	200.0000	200.0000	200.0000
MAST	*Ripser*	25100.0363	28515.8233	29119.0466
	PixHomology	245.0032	278.3272	284.2231

5 Conclusions

This study introduces the *PixHomology* algorithm, a novel approach for computing zero-dimensional PH on images. Using simplicial complexes constructed from pixel connections, our method offers a streamlined alternative to traditional approaches. Through an extensive comparative analysis with existing methods, our software demonstrates good execution time and a drastically reduced memory footprint. The efficiency achieved by *PixHomology* sets the stage for analyzing increasingly high-resolution images across diverse application domains. For example, in the realm of machine learning, PH was used for various tasks. In [4,13], for example, it showed great potential in segmenting tumor images while in [3] it is tested for segmenting astronomical images. Due to their typically large size, histological and astronomical images are commonly divided into smaller patches before processing. However, *PixHomology* enables the analysis of the entire image at once, enhancing the efficiency and reducing potential artifacts at the borders of each patch.

By enabling the processing of entire large-scale images, *PixHomology* paves the way for more comprehensive insights in various domains.

References

1. Agustsson, E, Timofte, R.: Ntire 2017 challenge on single image super-resolution: dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
2. Bauer, U.: Ripser: efficient computation of Vietoris-Rips persistence barcodes. J. Appl. Comput. Topol. **5**(3), 391–423 (2021)
3. Ceccaroni, R., Brutti, P., Castellano, M., Fontana, A., Merlin, E.: Topological persistence for astronomical image segmentation. In: The 51st Scientific Meeting of the Italian Statistical Society, pp. 1993–1998 (2022)
4. Chung, Y.M., Hu, C.S., Lawson, A., Smyth, C.: Topological approaches to skin disease image analysis. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 100–105 (2018)
5. Eaves, B.C.: Freudenthal triangulation F of G^n, Part I. In: A Course in Triangulations for Solving Equations with Deformations. Lecture Notes in Economics and Mathematical Systems, vol. 234, pp. 51–82. Springer, Berlin, Heidelberg (1984)
6. Morozov, D., et al.: Dionysus: a software library for topological data analysis (2012)
7. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. arXiv preprint arXiv:1907.07174 (2019)
8. Kaji, S., Sudo, T., Ahara, K.: Cubical ripser: software for computing persistent homology of image and volume data. arXiv preprint arXiv:2005.12692 (2020)
9. Kather, J., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. **16**, e1002730 (2019)
10. Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical report (2009)
11. LeCun, Y., Cortes, C., Burges, C.: MNIST handwritten digit database. ATT Labs (2010). http://yann.lecun.com/exdb/mnist
12. Munkres, J.: Elements of Algebraic Topology (1984)

13. Qaiser, T., Sirinukunwattana, K., Nakane, K., Tsang, Y.W., Epstein, D., Rajpoot, N.: Persistent homology for fast tumor segmentation in whole slide histology images. Procedia Comput. Sci. **90**, 119–124 (2016)
14. Wagner, H.: Slice, simplify and stitch: Topology-preserving simplification scheme for massive voxel data. In: 39th International Symposium on Computational Geometry (SoCG 2023). Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2023)
15. Wagner, H., Chen, C., Vuçini, E.: Efficient computation of persistent homology for cubical data (2012)

A Multi-Aspect Permutation Test for Selecting the Optimum Number of Clusters

Riccardo Ceccato(✉), Elena Barzizza, Nicoló Biasetton, and Marta Disegna

Department of Management and Engineering, University of Padova, Vicenza, Italy
riccardo.ceccato.1@unipd.it

Abstract. Cluster analysis is a multivariate statistical technique employed across various fields to detect clusters based on multivariate distance measures computed between objects. Being an exploratory analysis, the final outcome of a clustering algorithm heavily depends on several critical decisions made by the expert. A fundamental decision, widely discussed in the literature, especially for non-hierarchical clustering algorithms, concerns the identification of the final partition, i.e. the number of final clusters. In this paper, a permutation-based approach is introduced to reduce subjectivity in finding the final partition, assisting practitioners to tackle the specific challenge. In particular, an extension of the NonParametric Combination (NPC) methodology based on the idea of multi-aspect tests is adopted. In fact, the NPC method allows us to compare partitions using multiple clustering performance metrics, facilitating the identification of the best final partition. A simulation study, comprising four synthetic datasets, is presented and discussed to demonstrate the potentialities of the proposed approach.

Keywords: cluster analysis · permutation · NPC

1 Introduction

Cluster analysis is a statistical technique widely adopted to divide units into an appropriate set of clusters based on their distance. According to Tuena et al. (2020) [11], clustering is the task of grouping a set of objects in such a way that objects in the same group (i.e. a cluster) are more similar to each other than to those in other groups. Hard clustering algorithms are split into two main types, namely hierarchical and non-hierarchical. Among the non-hierarchical clustering algorithms, the K-means [8] and the Partitioning-Around-Medoids [7] are the most commonly adopted. One widely known problem in non-hierarchical clustering algorithms is that the number of final clusters in the sample data must be known in advance as it is required as input information to run the algorithm. Unfortunately, the number of clusters into which sample data drawn from a real case study best divides is usually unknown and this represents a real challenge both for researchers and practitioners [6].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 393–398, 2025.
https://doi.org/10.1007/978-3-031-64431-3_65

To this end, several different quality indices forclustering have been introduced, such as the separation index [5], the Dunn index [3], the Calinski-Harabasz index [2], and the Silhouette index [10]. Unfortunately, to the best of our knowledge, no criteria exist to determine which clustering index provides the best partition, which is usually selected subjectively by experts based on its usefulness from a practical and/or managerial point of view.

In this paper, a permutation-based approach is suggested to overcome this gap, providing a way to identify the best final partition without the need for subjective comparison of results from different clustering indices. In particular, the NonParametric Combination (NPC) methodology [9] is adopted for its multi-aspect nature that allows for the identification of the best partition in terms of both average quality (i.e. best trade-off between minimum internal homogeneity and maximum external separability) and cluster stability (i.e. variability).

The paper is structured as follows: Sect. 2 describes the suggested approach; in Sect. 3 the proposed procedure is applied to four synthetic datasets in order to highlight the approach's main features; in Sect. 4 we present our conclusions.

2 Method

Let us assume that we use the K-means algorithm to identify $C = 2$ different partitions P_c into k_c clusters that we want to compare in terms of V clustering quality indices.

The first part of the proposed approach requires us to consider a number of steps. Using $l = 1, \ldots, n$, and $c = 1, 2$, at each iteration l:

1. Randomly choose k_c initial cluster centers.
2. Apply the K-means algorithm and identify the best partition P_c into k_c clusters.
3. Compute V different clustering quality indices $\{Y_{clj}\}, j = 1, \ldots, V$.

This allows us to build two new datasets, $\mathbf{Y}_1 = \{Y_{1lj}, l = 1, \ldots, n, j = 1, \ldots, V\}$ and $\mathbf{Y}_2 = \{Y_{2lj}, l = 1, \ldots, n, j = 1, \ldots, V\}$, containing the information about the quality of the final partitions P_1 and P_2. In particular, each generic dataset is made up of n rows and V columns containing the values of the clustering quality indices for the clusters belonging to the two partitions. For each $j = 1, \ldots, V$ clustering quality index, test whether one partition is better than the other as follows:

$$\mu_{1j} \overset{<}{>} \mu_{2j},$$

where μ_{cj} is the mean of $\mathbf{Y}_{cj} = \{Y_{clj}, l = 1, \ldots, n\}, c = 1, 2$, and the direction of the comparison is determined by the nature of the index (i.e. if it needs to be maximized or minimized). These comparisons allow us to identify which partition on average provides the best quality indices.

In addition, for each $j = 1, \ldots, V$, the partition stability is evaluated as follows:

$$\sigma_{1j} < \sigma_{2j},$$

where σ_{cj} is the variance of $\mathbf{Y}_{cj}, c = 1, 2$. This in fact also allows us to evaluate how stable the partition is with different initializations of the algorithm.

Finally, we need to combine the indications provided by each individual test and to this end the NonParametric Combination methodology [9] can be applied.

Let $D_{lj} = Y_{1lj} - Y_{2lj}$ and $G_{lj} = |Y_{1lj} - \eta^{1j}| - |Y_{2lj} - \eta^{2j}|$, where Y_{clj} refers to the value of the j^{th} quality index for partition P_c at the l^{th} initialization of the K-means algorithm and η^{cj} is the median value of the j^{th} quality index for partition P_c.

Given that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are paired samples, according to this methodology, in the second part of the proposed approach we need to:

1. Compute the matrices $\mathbf{D}$ and $\mathbf{G}$.
2. Compute $T_{10j} = \sum_{l=1}^{n} D_{lj}$ and $T_{20j} = \sum_{l=1}^{n} G_{lj}$, $j = 1, \ldots, V$.
3. For $b = 1, \ldots, B$:
   (a) Consider a random permutation by randomly generating n signs S_l.
   (b) Compute the test statistics $T_{1bj} = \sum_{l=1}^{n} D_{lj} \cdot S_l^*$ and $T_{2bj} = \sum_{l=1}^{n} G_{lj} \cdot S_l^*$, $j = 1, \ldots, V$.
4. Compute a partial p-value and its simulated permutation distribution for each test statistic: $\boldsymbol{\lambda}_c = \{\lambda_{cbj}, b = 0, \ldots, B, j = 1, \ldots, V\}, c = 1, 2$.
5. Combine p-values by means of an appropriate combining function $\theta(\cdot)$ (e.g. Fisher's omnibus combining function [4]) to combine indications provided by each clustering quality index. A second-order combined test is therefore achieved as: $\mathbf{T}'_c = \theta(\boldsymbol{\lambda}_c), c = 1, 2$.
6. Compute the second-order p-values and their simulated permutation distribution: $\boldsymbol{\lambda}' = \{\lambda'_{cb}, c = 1, 2, b = 0, \ldots, B\}$.
7. Combine p-values by means of an appropriate combining function $\psi(\cdot)$. A third-order combined test is therefore achieved as: $\mathbf{T}'' = \psi(\boldsymbol{\lambda}')$.
8. Compute the global p-value for the third-order combined test to combine indications related to location and variability.

Further details about the algorithm used in the NonParametric Combination methodology and the calculation of the p-values and their permutation distributions are provided by Pesarin and Salmaso (2010) [9].

The global p-value achieved using this procedure can be used to choose between the partitioning P_1 and P_2 into k_1 and k_2 clusters, respectively.

However, we commonly wish to consider more than two possible partitions P_c (i.e. $C > 2$), with k_c generally varying between 2 and 10. In this case it is possible to integrate the aforementioned NPC-based approach into a ranking procedure by Arboretti et al. (2014) [1]. Therefore, we can:

1. Perform all possible pairwise comparisons between partitions P_c into $k_c \in \{2, \ldots, 10\}$ clusters in terms of V indices, using the proposed NPC-based test.
2. Compute a ranking of the partitions based on the adjusted (for multiplicity) p-values obtained through the $\frac{C \times (C-1)}{2}$ comparisons.
3. Select the partition into $\hat{K}$ clusters to which the first ranking position is assigned.

3 Preliminary Simulation Study

An exploratory study was conducted to evaluate the performances of the proposed procedure, using four synthetic datasets.

K represented the actual number of clusters, $N = 20$ the sample size, $H = 5$ the number of variables, $\boldsymbol{\gamma}_k = [\gamma_{k1}, \ldots, \gamma_{kJ}]$ the non-centrality parameters with $\boldsymbol{\gamma}_k = \boldsymbol{\gamma}_{k-1} + 2, k = 2, \ldots, K, \boldsymbol{\gamma}_1 = [5, \ldots, 5]$ and

$$\boldsymbol{\Sigma}_{H\times H} = \begin{bmatrix} 1 & 0.25 & \ldots & 0.25 \\ 0.25 & 1 & \ldots & 0.25 \\ \vdots & \vdots & \vdots & \vdots \\ 0.25 & 0.25 & \ldots & 1 \end{bmatrix}$$

the common scale matrix. We generated K samples as per the scenarios reported in Table 1. The R package `LaplacesDeamon` was used for data generation.

Table 1. Scenarios

Scenario	Distribution	K
Sc1	Multivariate normal	3
Sc2	Multivariate normal	5
Sc3	Multivariate power exponential, k=3	3
Sc4	Multivariate Student's t, df=4	3

Each scenario resulted in a synthetic dataset. On each dataset, we applied the K-means algorithm 1000 times (i.e. $n = 1000$), varying the number of clusters from 2 to 10. For each partition P_c, we measured the separation index, the Dunn index, and the Calinski-Harabasz index (i.e. $V = 3$). We then built the $n \times V$ dataset of quality indices on which we applied the proposed NPC-based procedure in order to retrieve the ranking of partitions. To evaluate the goodness of our proposal, we checked if the first ranking position was assigned to the partition with the right number of clusters (i.e. K). The R package `fpc` was used to compute cluster validation statistics, while the R code required to implement the NPC-based procedure is available upon request.

The achieved results are reported in Table 2. The interpretation of the achieved rankings is as follows: the lower the ranking position, the better the partition. It appears that in the four considered scenarios, the proposed procedure did indeed succeed in identifying the right number of clusters: 3 for Sc1, Sc3 and Sc4, and 5 for Sc2.

Table 2. Rankings

No. of clusters	Ranking Sc1	Ranking Sc2	Ranking Sc3	Ranking Sc4
2	2	4	7	5
3	**1**	5	**1**	**1**
4	9	3	3	3
5	3	**1**	2	7
6	3	2	7	9
7	3	6	7	3
8	8	8	4	8
9	7	7	5	2
10	3	9	5	5

4 Conclusions

In this paper we proposed a permutation-based procedure that exploits the Non-Parametric Combination methodology [9] and a ranking procedure [1] to help both researchers and practitioners choose the right number of final clusters when performing a cluster analysis. In particular, this approach allows experts to base their decision on multiple clustering quality indices and combine indications provided by each of them. The proposed ranking procedure aims to assign the top rank to the partition with the optimal number of clusters, i.e. the partition that significantly outperforms the others in terms of both the average and the dispersion of a set of clustering quality indices.

Four different synthetic datasets were presented to illustrate the usefulness and effectiveness of the suggested procedure. Results show that the method is promising in terms of helping experts make the challenging decision about the number of clusters. In this preliminary study, the separation, the Dunn, and the Calinski-Harabasz indices were used as clustering quality indices. An extensive simulation study and applications to real case-studies are planned to better evaluate the performance of the suggested method and to evaluate the inclusion of other clustering quality indices, such as the Silhouette index.

Acknowledgement. Nicoló Biasetton and Marta Disegna thank MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-Generation EU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). CUP MICS C93C220052800.

Elena Barzizza and Riccardo Ceccato thank MOST—Sustainable Mobility National Research Centre and received funding from the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR)—Missione 4 Compo.nente 2, Investimento 1.4—D.D. 1033 17 June 2022, CN00000023).

References

1. Arboretti, R., Bonnini, S., Corain, L., Salmaso, L.: A permutation approach for ranking of multivariate populations. J. Multivar. Anal. **132**, 39–57 (2014)
2. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. Theory Methods **3**(1), 1–27 (1974)
3. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. J. Cybern. **4**(1), 95–104 (1974)
4. Fisher, R.A.: Statistical methods for research workers. In: Breakthroughs in statistics, pp. 66–70. Springer (1992)
5. Hennig, C., Liao, T.F.: How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J. R. Stat. Soc. Ser. C Appl. Stat. **62**(3), 309–369 (2013)
6. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
7. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley (2005)
8. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Oakland, CA, USA (1967)
9. Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. Wiley (2010)
10. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
11. Tuena, C., Chiappini, M., Repetto, C., Riva, G.: Artificial intelligence in clinical psychology. In: Reference Module in Neuroscience and Biobehavioral Psychology. Elsevier (2020)

Multilevel Bivariate Areal Modelling for School Data: An Application with R-INLA

Leonardo Cefalo[1(✉)] and Virgilio Gómez-Rubio[2]

[1] Università degli studi di Bari Aldo Moro, Bari, Italy
leonardo.cefalo@uniba.it
[2] Universitad de Castilla-La Mancha, Albacete, Spain

Abstract. This paper studies the relationship between the territorial endowment of school infrastructure and services in Italy and the performance of high school students, expressed by the Invalsi scores for the second year of high school, by means of spatial modelling. We fit a multilevel Bayesian regression model with a latent spatial effect modelled as a conditional autoregressive (CAR) process. Since the Invalsi scores regard two subjects (Italian and Mathematics) we fit a bivariate model, in which the relationship between the two subjects is accounted for by the random effect. The computation follows the method of the Integrated Nested Laplace Approximation (INLA), implemented in the `R-INLA` software.

Keywords: Areal models · multivariate models · INLA · Invalsi

1 Introduction

The study of territorial disparities in infrastructure and services endowment often relies on the methodologies of spatial statistics. In the present context we make use of the municipality-level Invalsi scores as a standardised and spatially homogeneous indicator of the results obtained by school students and explore the relationship between these scores, reported for the second year of high school in the subjects of Italian and Mathematics, and the infrastructural state of municipalities expressed in terms of centrality degree, the availability of ultra-broadband internet connection in schools, and the school accessibility by means of urban public transport.

To take into account the territorial dimension of the data we introduce a bivariate spatially structured latent effect defined at a higher level, either for provinces or for catchment areas of the infrastructural poles, and treated as a Conditional Autoregressive (CAR) Process. The model is computed through the Integrated Nested Laplace Approximation INLA, [11], which is implemented in a specific and self-sufficient `R` package https://www.r-inla.org/home. The multivariate extension of the CAR model we employ is already implemented in `R` [7].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 399–405, 2025.
https://doi.org/10.1007/978-3-031-64431-3_66

2 Data

We collect data from 874 municipalities, referring to school year 2022/2023.

Invalsi scores The Italian Institute for the Evaluation of the Education System (Invalsi) is a research institute whose primary mission is to assess the quality of the public education system through a censuary survey regarding the 2nd and 5th year of primary school, the last year of middle school, 2nd and 5th year of high school. The strong territorial disparity in high school scores motivates us to focus on the second year of high school, being it part of the compulsory education cycle (differently from the last year). Scores are expressed via a specific psychometric model having by construction a global mean of 200 points and a school-level standard deviation of 40 points.

Covariates. The choice of covariates derives from forward selection. To the state of our findings, the most eligible candidates are the following:

Urban Public Transport. The municipality-level percentage of high schools served by urban public transport, reported in the School Buildings section of the Unique School Data Portal [6].

Ultra-Broadband Activation Status. The municipality-level percentage of high schools provided with ultra-broadband at the beginning of school year 2022/23. Ultra-broadband is defined as internet connection with a maximum speed of 1 gigabit/second and a minimum guaranteed speed of 100 megabits/second.

Inner Areas. The classification of inner areas is published by the Italian National Institute of Statistics [5] and includes six classes; classes A and B identify the poles and the intermunicipality poles, namely the municipalities (A) or the clusters of contiguous municipalities (B) endowed with fundamental school, transport and health infrastructure. Classes from C to F are ranked in terms of travel time to the closest pole, referred to as the destination pole. We consider central the municipalities in classes A and B (serving as destination poles), intermediate the municipalities in classes C and D, and peripheral the municipalities in classes E and F.

3 Methods

The Integrated Nested Laplace Approximation. The INLA is a computational method, one of whose main fields of application is spatial analysis see e.g. [1]. The baseline is a generic hierarchical regression model, where $\mathbf{y}$ is the response variable; θ is a set of parameters and latent effects, Ψ is the set of hyperparameters, whose size is usually small (in the order of units).

$\theta|\Psi$ is assumed to be a Gaussian Markov Random Field see [10] with joint distribution $p(\theta \mid \Psi) = N(\mathbf{0}, (\mathbf{Q}(\Psi))^{-1})$. Single observations are considered independent conditioned on the parameters, i.e. $p(\mathbf{y}|\theta, \Psi) = \prod_{i=1}^{N} p(y_i \mid \theta, \Psi)$. The

distribution $\theta|\mathbf{y}$ is obtained by integrating numerically the full posterior $\theta|\mathbf{y}, \Psi$ over Ψ, which is feasible provided that Ψ has a low dimensionality. The INLA basically consists in approximating the full conditional distribution via the Taylor expansion of $p(\theta \mid \Psi)p(\mathbf{y} \mid \theta, \Psi)$ truncated at the second order and located at the mode of $\theta|\Psi$.

The CAR Model. At the unit level, the Gaussian Conditionally Autoregressive (CAR) or autonormal process z_i is defined through the conditional distribution, which is Normal with mean $E[z_i|\mathbf{z_{-i}}] = \sum_{j \neq i} \frac{w_{ij}}{n_i} z_j$ and variance $\frac{\sigma^2}{n_i}$ [2]. The coefficient w_{ij} is a generic element of the neighbourhood matrix $\mathbf{W}$, equalling 1 if sites i and j are neighbours and 0 otherwise; $n_i := \sum_{j \neq i} w_{ij}$ represents the number of neighbours of site i and the i-th row sum of $\mathbf{W}$; σ^2 is a global scale parameter. The joint distribution is still Gaussian with mean $\mathbf{0}$ and precision $\sigma^{-2}(\mathbf{D} - \mathbf{W})$, being $\mathbf{D}$ the diagonal matrix of the neighbours cardinality, i.e. $D_{11} = n_1, D_{22} = n_2 ...$ and so on.

This distribution is evidently improper in that the precision element $(D-W)$ has all rows summing to zero and is therefore singular. An alternative model would employ an additional hyperparameter α such that the joint distribution becomes proper, by defining $\sigma^{-2}(\mathbf{D}-\alpha\mathbf{W})$ as the precision matrix. Higher values of α imply stronger spatial dependence, the limit case with $\alpha = 1$ being the improper CAR.

This model can be generalised to the case of p variables [3] and is implemented in `R-INLA` for a number of models (see [7,8]. To our aims, referring respectively to the improper and proper model as IMCAR and PMCAR, it holds that:

$$\text{IMCAR:} \quad \mathbf{z} \sim N\left(\mathbf{0}, \Lambda^{-1} \otimes (\mathbf{D} - \mathbf{W})^{-1}\right) \tag{1}$$

$$\text{PMCAR:} \quad \mathbf{z} \sim N\left(\mathbf{0}, \Lambda^{-1} \otimes (\mathbf{D} - \alpha\mathbf{W})^{-1}\right) \tag{2}$$

The global precision parameter Λ is defined as

$$\Lambda = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1p}\sigma_1\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p}\sigma_1\sigma_p & \rho_{2p}\sigma_2\sigma_p & \dots & \sigma_p^2 \end{pmatrix}^{-1}$$

Where ρ terms denote the correlations among random effects.For the independent model the correlations are zero and Λ is reduced to its diagonal.

4 Model

As a preliminary step, we center the observed Invalsi scores around zero. Scores in subjects 1 and 2 for municipality j in macro-area i are thus modelled:

$$\begin{aligned} \begin{pmatrix} y_{i,j,1} \\ y_{i,j,2} \end{pmatrix} = \begin{pmatrix} \beta_{0,1} \\ \beta_{0,2} \end{pmatrix} + x_{i,j,1} \begin{pmatrix} \beta_{1,1} \\ \beta_{1,2} \end{pmatrix} + x_{i,j,2} \begin{pmatrix} \beta_{2,1} \\ \beta_{2,2} \end{pmatrix} + x_{i,j,3} \begin{pmatrix} \beta_{3,1} \\ \beta_{3,2} \end{pmatrix} + \\ + x_{i,j,4} \begin{pmatrix} \beta_{4,1} \\ \beta_{4,2} \end{pmatrix} + \begin{pmatrix} z_{i,1} \\ z_{i,2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,j,1} \\ \varepsilon_{i,j,2} \end{pmatrix} \end{aligned} \tag{3}$$

Being ε two independent IID white noise processes with mean zero and variances ω_1 and ω_2.

Given the unbounded domain of $\mathbf{y}$ we firstly model it as a Gaussian process; to deal with high dispersion we also fit a t-Student model and compare the two models.

Fixed effects have independent Gaussian priors with mean zero and variance = 1000. White noise precisions have Gamma priors with shape = 10^{-3} and rate = 10^{-3}.

Random effects are defined at the level of provinces (105 areas) and catchment areas of central municipalities (206 areas) and modelled as bivariate zero-mean PMCAR or IMCAR. To solve the identifiability issue in the IMCAR, we impose a link among the connected components of the spatial domain (i.e. between islands and mainland) and set a unique sum-to-zero constraint.

The prior for the dense Λ precision parameter is the Wishart distribution with 2 degrees of freedom and scale parameter equal to the identity matrix 5as suggested in [7], i.e.: $\Lambda \sim Wishart_2(I_2)$, while for the diagonal Λ in the independent model the square roots of the entries have a uniform prior defined on $]0; \infty[$.

α has a uniform prior on $[-1, 1]$.

5 Results

To compare the models, we compute the logarithmic sum of the Conditional Predictive Ordinates [9], a cross-validatory diagnostic defined as $-\sum_i ln\left(p(y_i|y_{-i})\right)$, the Deviance Information Criterion (DIC), and the Mean Square Error (MSE) of the posterior means of fitted values. The dependent CAR is preferable to the independent one. Using a finer support for the latent effects, i.e. the 206 pole catchment areas rather than the 105 provinces, improves the fitting but the province-level model is preferable in terms of DIC and CPO. As an example, we show the summary of fixed effects and hyperparameters for the PMCAR

Table 1. Model diagnostics

			Independent models			Dependent models		
z mod.	z lev.	y dens.	CPO	DIC	MSE	CPO	DIC	MSE
ICAR	Prov	Norm	6743.611	13486.16	239.534	6712.652	13425.18	235.395
ICAR	Pole	Norm	6751.424	13501.53	238.130	6716.743	13433.49	232.735
PCAR	Prov	Norm	6743.799	13486.86	238.898	6713.650	13427.26	233.933
PCAR	Pole	Norm	6752.137	13502.82	237.619	6718.280	13436.99	230.576
ICAR	Prov	t-St	6697.211	13397.33	240.518	6663.304	13329.06	237.822
ICAR	Pole	t-St	6708.396	13418.51	238.503	6666.699	13335.29	233.927
PCAR	Prov	t-St	6697.728	13398.00	240.024	6664.105	13331.08	236.785
PCAR	Pole	t-St	6708.725	13420.03	238.038	6667.732	13337.97	232.761

Table 2. fixed effects and PCAR hyperparameters

		Gaussian				t-Student			
Subj	Effect	Mean	SD	LB	UB	Mean	SD	LB	UB
MAT	Intercept	−0.897	5.172	−11.691	10.135	−0.700	5.380	−11.831	10.54
ITA	Intercept	−2.722	4.494	−12.151	6.891	−0−717	4.649	−10.372	8.993
MAT	Central	2.769	0.910	0.984	4.555	2.713	0.807	1.132	4.296
ITA	Central	4.131	0.989	2.191	6.071	3.392	0.827	1.776	5.021
MAT	Peripheral	−2.306	1.004	−4.275	−0.336	−2.208	0.926	−4.024	−0.390
ITA	Peripheral	−1.442	1.082	−3.562	0.680	−2.113	0.990	−4.049	−0.165
MAT	BB Active	3.233	1.078	1.126	5.340	2.841	1.028	0.828	4.858
ITA	BB Active	3.419	1.154	1.155	5.682	2.713	1.067	0.623	4.810
MAT	Urban tpt	2.456	1.055	0.391	4.522	2.418	1.009	0.439	4.399
ITA	Urban tpt	3.396	1.109	1.222	5.573	2.413	1.035	0.388	4.448
Subj	Hyperpar	Mean	SD	LB	UB	Mean	SD	LB	UB
	α	0.992	0.0086	0.969	0.999	0.993	0.0062	0.976	0.999
MAT	σ^2	44.761	12.207	24.899	72.516	45.216	10.719	27.427	69.321
ITA	σ^2	33.206	9.470	18.549	55.491	31.753	8.054	18.683	50.147
	ρ	0.942	0.0505	0.804	0.993	0.954	0.035	0.859	0.991
MAT	ω	111.043	5.421	100.783	122.069	113.105	8.383	97.523	130.431
ITA	ω	133.041	6.461	120.91	146.279	140.275	12.843	116.84	167.248
MAT	ν					5.910	1.280	4.001	8.980
ITA	ν					4.546	0.789	3.331	6.406

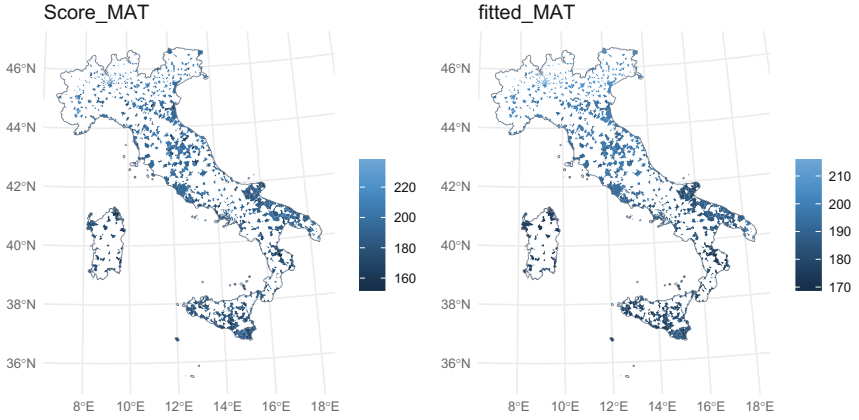


Fig. 1. Observed and predicted Mathematics scores

with province-level random effects in Table 2. All covariates range in $[0, 1]$ being them either Boolean or proportions. Credible intervals boundaries are the 2.5th and 97.5th percentile; hyperparameter ν denotes the degrees of freedom in the

t-Student distribution. The regularising factor α ranges close to one, denoting a strong spatial association among the random effects. The high value of ρ confirms that dependent models are more adequate than independent ones. In Fig. 1 we display the fitting of Mathematics scores through the T-Student proper model with province-level latent effects.

6 Concluding Remarks

The infrastructural state of schools and municipalities appears to have a significant impact on students performance. Our preliminary findings suggest that indeed the peripheral areas have a disadvantage in education quality. Other than between-municipalities factors, however, we find that the Invalsi scores are also driven by a between-provinces spatial trend, confirming in turn the significance of the territorial divide shaping many aspects of Italian society.

The first and most immediate future development would be generalising this analysis to different school grades and different years; also different likelihoods are worthy of being tested for these data. Finally, an aspect worthy of investigation may be the study of model constraints aimed at estimating the random effects as orthogonal complements to the fixed effects [4].

Overall, the mapping of infrastructure access and social vulnerability requires adequate statistical computational methods, and `R-INLA` proves to be a flexible tool to this aim.

Acknowledgement. This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUPH93C22000650001). 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. Bakka, H., Rue, H., Fuglstad, G.A., et al.: Spatial modeling with R-INLA: a review. WIREs Comput. Stat. **10**, e1443 (2018). https://doi.org/10.1002/wics.1443
2. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. Ser. B **36**(2), 192–236 (1974). https://www.jstor.org/stable/2984812
3. Gelfand, A.E., Vounatsou, P.: Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics **4**(1), 11–25 (2003). https://doi.org/10.1093/biostatistics/4.1.11
4. Hodges, JS., Reich, BJ.: Adding spatially-correlated errors can mess up the fixed effect you love. Am. Stat **64**(4), 325–334 (2010). https://doi.org/10.1198/tast.2010.10052
5. ISTAT: La geografia delle aree interne nel 2020: vasti territori tra potenzialità e debolezze (2020). https://www.istat.it/it/files//2022/07/FOCUS-AREE-INTERNE-2021.pdf
6. Ministry of Education, University and Research: Portale Unico dei Dati sulla scuola. https://dati.istruzione.it/opendata/

7. Palmí-Perales, F., Gómez-Rubio, V., Martinez-Beneito, M.A.: Bayesian multivariate spatial models for lattice data with INLA. J. Stat. Softw. **98**(2), 1–29 (2021). https://doi.org/10.18637/jss.v098.i02
8. Palmí-Perales, F., Gómez-Rubio, V., Bivand, R. S., Cameletti, M., Rue, H.: Bayesian inference for multivariate spatial models with INLA. R J. **15/3** (2023)
9. Pettit, L. I.: The conditional predictive ordinate for the normal distribution. J. Roy. Stat. Soc. Ser. B **52**(1), 175–184 (1990). https://www.jstor.org/stable/2345658
10. Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. 1st edn. Chapman and Hall/CRC (2005). https://doi.org/10.1201/9780203492024
11. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B **71**(2), 319–392 (2009). https://doi.org/10.1111/j.1467-9868.2008.00700.x

RODESSA for Robust Multivariate Singular Spectrum Analysis

Fabio Centofanti[1(✉)], Mia Hubert[2], Biagio Palumbo[1], and Peter J. Rousseeuw[2]

[1] Department of Industrial Engineering, University of Naples Federico II, Napoli, Italy
fabio.centofanti@unina.it

[2] Section of Statistics and Data Science Department of Mathematics,, KU Leuven, Belgium

Abstract. Multivariate Singular Spectrum Analysis (MSSA) is a powerful and widely used nonparametric method for multivariate time series. However, MSSA lacks robustness against outliers because it relies on the singular value decomposition, which is very sensitive to the presence of anomalous values. MSSA can then give biased results and lead to erroneous conclusions. In this paper, we present the method proposed by [2], named *RObust Diagonalwise Estimation of SSA* (RODESSA), which is robust against the presence of cellwise and casewise outliers. A real data example about temperature analysis in passenger railway vehicles demonstrates the practical utility of the proposed approach.

Keywords: Casewise outliers · Cellwise outliers · Iteratively reweighted least squares · Multivariate time series · Robust statistics

1 Introduction

Time series analysis is essential for comprehending and forecasting the patterns within sequential data across various fields. Given that real-world phenomena frequently encompass multiple interconnected variables, numerous multivariate time series methodologies have been developed to address this complexity [8]. Singular spectrum analysis (SSA) has emerged as a powerful non-parametric tool for extracting valuable insights from time-dependent data [3,4]. Its multivariate version is abbreviated as MSSA. We give a brief description of its common version. Consider a p-variate time series $\mathbb{X} = \left(\mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(p)}\right)$, i.e., a collection $\{\mathbb{X}^{(j)} = (x_i^{(j)})_{i=1}^{N}, \, j = 1, \ldots, p\}$ of p time series of length N. Multivariate SSA then proceeds by the following four successive steps.

1. Embedding. In the embedding step, the multivariate time series $\mathbb{X}$ is turned into a big trajectory matrix $\boldsymbol{X}$. Let L be an integer called *window length*, $1 < L < N$. For each time series $\mathbb{X}^{(j)}$ we then form $K_u = N - L + 1$ lagged vectors $\boldsymbol{X}_i^{(j)} = \left(x_i^{(j)}, \ldots, x_{i+L-1}^{(j)}\right)^T$ for $1 \leqslant i \leqslant K_u$. The trajectory matrix of $\mathbb{X}$ is a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 406–411, 2025.
https://doi.org/10.1007/978-3-031-64431-3_67

matrix of size $L \times K$ with $K = pK_u$, and has the form

$$\boldsymbol{X} = \left[\boldsymbol{X}_1^{(1)} : \ldots : \boldsymbol{X}_{K_u}^{(1)} : \ldots : \boldsymbol{X}_1^{(p)} : \ldots : \boldsymbol{X}_{K_u}^{(p)}\right] = \left[\boldsymbol{X}^{(1)} : \ldots : \boldsymbol{X}^{(p)}\right]$$

where $\boldsymbol{X}^{(j)} = \left[\boldsymbol{X}_1^{(j)} : \ldots : \boldsymbol{X}_{K_u}^{(j)}\right]$ is the trajectory matrix of the univariate series $\mathbb{X}^{(j)}$. Note the (anti-)diagonal structure of each $\boldsymbol{X}^{(j)}$, making it a so-called *Hankel matrix*. The entire matrix $\boldsymbol{X}$ is thus a *stacked Hankel matrix*.

2. Decomposition. This step performs the SVD of the trajectory matrix $\boldsymbol{X}$, yielding

$$\boldsymbol{X} = \sum_{r=1}^{d} \beta_r \boldsymbol{u}_r \boldsymbol{v}_r^T \tag{1}$$

where $d = \operatorname{rank}(\boldsymbol{X})$, and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ are the left and right singular vectors. The ordered singular values are $\beta_1 \geqslant \cdots \geqslant \beta_d > 0$, and the matrices $\boldsymbol{X}_r = \beta_r \boldsymbol{u}_r \boldsymbol{v}_r^T$ all have rank 1. The triple $(\beta_r, \boldsymbol{u}_r, \boldsymbol{v}_r)$ is called the r-th eigentriple of the matrix $\boldsymbol{X}$.

3. Grouping. The grouping step corresponds to splitting the terms of (1) into several disjoint groups and summing the matrices within each group. We focus on the case where two main groups are created, i.e.,

$$\boldsymbol{X} = \widehat{\boldsymbol{X}}_q + \boldsymbol{R} \tag{2}$$

where the estimate of the trajectory matrix $\widehat{\boldsymbol{X}}_q$ corresponding to the signal is a sum like (1) but only for the first q eigentriples, whereas the residual matrix $\boldsymbol{R} = \boldsymbol{X} - \widehat{\boldsymbol{X}}_q$ is associated with the noise. We can see $\widehat{\boldsymbol{X}}_q$ as a low-rank approximation of rank q of the trajectory matrix.

4. Reconstruction. In this step, the fitted matrix $\widehat{\boldsymbol{X}}_q$ is transformed back to the form of the time series $\mathbb{X}$. In each submatrix $\widehat{\boldsymbol{X}}_q^{(j)}$ we compute average entries as follows. Denote an (anti-)diagonal as $A_i = \{(l,k) : l+k = i+1, 1 \leqslant l \leqslant L, 1 \leqslant k \leqslant K_u\}$ with its cardinality $n_i = |A_i|$. Each matrix $\widehat{\boldsymbol{X}}_q^{(j)} = (\widehat{X}_{lk}^{(j)})_{lk}$ is turned into a new series $\widehat{\mathbb{X}}^{(j)} = (\hat{x}_i^{(j)})_{i=1}^N$ of length N, with $\hat{x}_i^{(j)} := \frac{1}{n_i} \sum_{(l,k)\in A_i} \widehat{X}_{lk}^{(j)}$. The reconstructed multivariate time series is then given by $\widehat{\mathbb{X}} = \left(\widehat{\mathbb{X}}^{(1)}, \ldots, \widehat{\mathbb{X}}^{(p)}\right)$.

Multivariate time series data may exhibit outliers due to various factors such as measurement errors, data entry mistakes, sensor malfunctions, or unexpected events. These outliers can introduce bias in statistical measures, affect parameter estimation, and lead to inaccuracies in forecasting, potentially resulting in flawed decision-making. As in the multivariate setting [7], we distinguish between a *cellwise outlier*, which is an outlying value $x_i^{(j)}$ in the j-th univariate time series only, and a *casewise outlier*, where at a timepoint i several or all of the values $x_i^{(j)}$ deviate simultaneously. Classical SSA and MSSA are particularly sensitive to outliers due to their reliance on SVD, which is a non-robust least squares procedure. Robust methods for time series analysis have predominantly focused on classical parametric approaches. However, none of these low-rank

approximations take the special diagonal structure of the trajectory matrix into account. The new MSSA approach called *RObust Diagonalwise Estimation of SSA* (RODESSA) is tailored to this situation.

2 Methodology

RODESSA is unique in that it explicitly considers the occurrence of outliers and their residuals within the stacked Hankel structure of the trajectory matrix. It is meaningful to talk about anomalous diagonals rather than anomalous rows of the trajectory matrix $\boldsymbol{X}$. Therefore RODESSA approximates the trajectory matrix $\boldsymbol{X}$ by $\widehat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{V}^T$ obtained by minimizing

$$L_{\rho_1,\rho_2}\left(\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T\right) := \sum_{i=1}^{N} p n_i \hat{\sigma}_2^2 \rho_2 \left(\frac{1}{\hat{\sigma}_2} \sqrt{\frac{1}{p}\sum_{j=1}^{p} \hat{\sigma}_{1,j}^2 \rho_1\left(\frac{r_i^{(j)}}{\hat{\sigma}_{1,j}}\right)} \right) \tag{3}$$

over $(\boldsymbol{U}, \boldsymbol{V})$ with $\operatorname{rank}(\boldsymbol{U}\boldsymbol{V}^T) = q$. The $r_i^{(j)}$ are the norms of the diagonal residuals of series j, given by $r_i^{(j)} := \sqrt{\frac{1}{n_i}\sum_{a=1}^{n_i}\left(x_i^{(j)} - \hat{x}_{ia}^{(j)}\right)^2}$ where $\hat{x}_{ia}^{(j)}$ are the predicted cells. The fixed scales $\hat{\sigma}_{1,j}$ standardize the $r_i^{(j)}$, and the fixed overall scale $\hat{\sigma}_2$ standardizes the quantities $r_i := \sqrt{\frac{1}{p}\sum_{j=1}^{p} \hat{\sigma}_{1,j}^2 \rho_1\left(\frac{r_i^{(j)}}{\hat{\sigma}_{1,j}}\right)}$ based on all p coordinates. In our implementation ρ_1 and ρ_2 are biweight functions [6]. The goal of the normalization by n_i and by $\hat{\sigma}_{1,j}$ inside ρ_1 in (3) is to give the $r_i^{(j)}/\hat{\sigma}_{1,j}$ similar average sizes for each i and j. Given initial estimates $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$, the scale estimates $\hat{\sigma}_{1,j}$ and $\hat{\sigma}_2$ are computed as M-scales of the quantities $r_i^{(j)}$ and the r_i, all with respect to the fit $\boldsymbol{U}_0\boldsymbol{V}_0^T$.

The statistical meaning of the objective function in (3) can be better understood by connecting it to the definition of cellwise and casewise outliers. Indeed, cellwise outliers correspond to anomalous diagonals whereas casewise outliers simultaneously affect several diagonals corresponding to the same timepoint. A cellwise outlier results in large absolute values of $x_i^{(j)} - \hat{x}_{ia}^{(j)}$ in the i-th diagonal of the j-th time series. The idea of RODESSA is that these outliers could be better identified by simultaneously considering all $x_i^{(j)} - \hat{x}_{ia}^{(j)}$ in the same diagonal as a measure of cellwise outlyingness. In this sense, $r_i^{(j)}$ measures how prone the i-th value $x_i^{(j)}$ of time series j is to be a cellwise outlier, and the function ρ_1 reduces the effect of cellwise outliers on the final estimates, because large values of $r_i^{(j)}$ contribute less to the objective function. Similarly, a casewise outlier results in large values of $r_i^{(j)}$ for many j, therefore the diagonals corresponding to timepoint i of the time series are large. To take advantage of how the stacked Hankel structure of the trajectory matrix reacts to the presence of casewise outliers, the RODESSA objective function simultaneously considers all $r_i^{(j)}$ corresponding to a given i as a measure of casewise outlyingness. In this sense, r_i measures how

prone the multivariate $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(p)})^T$ is to be a casewise outlier, and the function ρ_2 reduces the effect of casewise outliers on the final estimates, because large values of r_i contribute less to the objective function. Additional details on the algorithm, implementation, and forecasting method of RODESSA can be found in [2].

3 A Real Data Example: Temperature analysis in Passenger Railway Vehicles

In this section, a real data example from the railway industry illustrates the applicability and potential of RODESSA. Recently, railway transportation in Europe has become a competitive option, prompting operators to focus on improving passenger satisfaction. A key challenge is ensuring optimal thermal comfort within passenger rail coaches. To tackle this, new European standards like UNI EN 14750 [1] have been introduced, regulating factors like air temperature, humidity, and air quality. As a result, railway companies are actively installing sensors to gather data from on-board HVAC systems, crucial for maintaining passenger comfort. This data is used to enhance the performance of HVAC systems, ultimately benefiting passenger experience. We analyze operational data from HVAC systems installed on a passenger 6-coach train operating during the summer season [5]. Specifically, the inside temperature of each coach was recorded about every four minutes from 10:00 to 22:00, yielding $N = 176$ observations of a multivariate time series with $p = 6$ components. RODESSA is applied in its default form. Figure 1 shows the enhanced time series plot that is a new graphical representation to assist with outlier detection [2]. The vertical dashed lines in Fig. 1 denote casewise outliers. It appears that the temperature inside the six coaches deviates significantly from the expected pattern in an almost periodic manner. This deviation is especially pronounced during the period between 17:00 and 18:00, where discussions with domain experts revealed that measurements were taken during train stops at terminal stations. During these stops, coach doors were left open, leading to an increase in temperature inside the coaches, particularly notable due to the summer season. Consequently, these measurements do not accurately represent the typical operating conditions of the train and are not suitable for characterizing temperature dynamics. Furthermore, the reconstructed time series illustrates a periodic increase and decrease in temperature inside the coaches, aligning with the on/off control of the HVAC system. It's worth noting that not all casewise outliers are identified as cellwise outliers. For example, the casewise outlier shortly after 16:00 doesn't have any component flagged as a cellwise outlier, but temperatures were relatively low in all six coaches simultaneously. The ability to detect such nuances is a key feature of the proposed method.

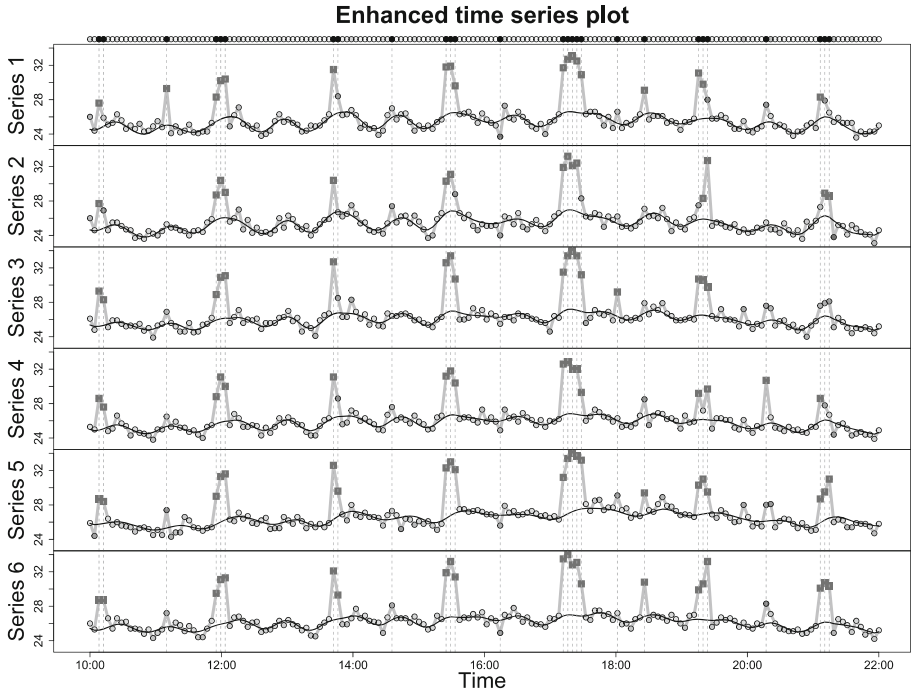


Fig. 1. Enhanced time series plot from applying RODESSA to the multivariate time series of temperatures. The solid lines correspond to the reconstructed multivariate time series. The original cells are connected by solid lines. Cells flagged as cellwise outliers are shown as solid squares. Casewise outliers are indicated by a solid black circle at the top of the plot and a vertical dashed gray line.

4 Conclusions

Multivariate singular spectrum analysis is a technique for fitting vector time series and making forecasts. A cell of the time series, that is, the value of one of its coordinates at a given timepoint, corresponds to a diagonal in the corresponding submatrix. A case of the time series, that is, all of the coordinates at a given time point, corresponds to diagonals in all of the submatrices. Sometimes cells are outlying, and sometimes entire cases. The classical MSSA method can be strongly affected by outliers of either type, because it is based on a low-rank singular value. Robust methods, which have been proposed in the literature, do not take the diagonal structure into account, and even a small number of outliers can create a large number of outlying diagonal entries that affect many rows and columns of the matrix. RODESSA tries to resolve this issue by explicitly taking the diagonal structure into account when decomposing the matrix. We have applied the RODESSA method to a real multivariate time series about temperatures in passenger railway vehicles. This illustrates its good reconstructing

performance, as well as the convenient outlier detection by the enhanced time series plot.

Acknowledgements. The contribution of F. Centofanti has been financially supported by Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 5 Componente 2, Investimento 1.3-D.D. 1551.11-10-2022, PE00000004 within the Extended Partnership MICS (Made in Italy - Circular and Sustainable). The activity of B. Palumbo is supported by the MOST - Sustain- able Mobility National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 - D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. ed responsible for them.

References

1. British Standards Institution: BS-EN 14750: railway applications - air conditioning for urban and suburban rolling stock - part 1: comfort parameters (2006)
2. Centofanti, F., Hubert, M., Palumbo, B., Rousseeuw, P.J.: Multivariate singular spectrum analysis by robust diagonalwise low-rank approximation. arXiv preprint arXiv:2310.01182 (2023)
3. Golyandina, N., Korobeynikov, A., Zhigljavsky, A.: Image processing. In: Singular Spectrum Analysis with R. UR, pp. 231–270. Springer, Heidelberg (2018). https://doi.org/10.1007/978-3-662-57380-8_5
4. Golyandina, N., Zhigljavsky, A.: SSA for forecasting, interpolation, filtering and estimation. In: Singular Spectrum Analysis for Time Series. SS, pp. 91–146. Springer, Heidelberg (2020). https://doi.org/10.1007/978-3-662-62436-4_3
5. Lepore, A., Palumbo, B., Sposito, G.: Neural network based control charting for multiple stream processes with an application to HVAC systems in passenger railway vehicles. Appl. Stoch. Model. Bus. Ind. **38**(5), 862–883 (2022)
6. Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust Statistics: Theory and Methods (with R). Wiley (2019)
7. Raymaekers, J., Rousseeuw, P.J.: Challenges of cellwise outliers. In: Econometrics and Statistics (2024)
8. Tsay, R.S.: Multivariate Time Series Analysis: with R and Financial Applications. Wiley (2013)



Sequence Analysis for Vaccination Data

Chiara Chiavenna[1,2(✉)], Danilo Bolano[1,3], and Alessia Melegaro[1,2]

[1] DONDENA Centre for Research on Social Dynamics and Public Policy, Bocconi University, Milan, Italy
chiara.chiavenna@gmail.com
[2] BIDSA Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy
[3] Department of Statistics, Informatics and Applications, University of Florence, Florence, Italy

Abstract. This work introduces and discusses the use of state sequence analysis (SSA) as an analytic framework for vaccination data. SSA is a well-known data-mining tool used in sociological literature however, to date, epidemiological studies featuring longitudinal categorical data have mostly relied on probabilistic models or summary measures at the aggregate level. Leveraging data on the uptake of hexavalent vaccine in the Lombardy region, we show how SSA helps in disentangling timeliness of vaccination.

Keywords: sequential data · health trajectories · data mining

1 Introduction

When examining the occurrence of events throughout an individual's life, it is of interest to analyze the resulting so-called life trajectory, particularly focusing on the timing of crucial events. Such data is often gathered through large-scale surveys, collecting individual-level information in a prospective or retrospective manner, or through administrative registers such as electronic health records (EHR). While extensive research in demography and sociology has worked towards reconstructing life course trajectories [1,2], less attention has been paid to the unfolding of healthcare accesses. Specifically, to the best of our knowledge, this is the first study using individual-level EHR to accurately describe the timing and timeliness of childhood immunization.

The literature on vaccination timeliness is currently organized along two lines. First, studies that describe uptake in the form of coverage, i.e. proportion of individuals who either completed or initiated a vaccination course by a given birthday, only assess timeliness with respect to a fixed age [3–6]. Alternatively, when delay with respect to an ideal timing is measured as a continuous outcome, arbitrary thresholds are used to define whether vaccines were received 'late' or 'on time', the most common one being a 28-day window [7,8]. Two key limitations

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 412–418, 2025.
https://doi.org/10.1007/978-3-031-64431-3_68

affect both methods: firstly, single-dose uptakes are considered as independent—rather than interrelated—events; secondly, time-detailed data are reduced to a dichotomous outcome, with a resulting loss of information.

We provide two contributions to the relevant literature. We present a novel comprehensive analysis of childhood immunization patterns using EHR of a population of around 10 million residents. More importantly, in so doing, we shift the focus from measuring the complete vaccine uptake towards describing the vaccination path as a combination of interdependent events where not only timing matters, but also duration between events.

2 The Study of the Life Course of an Individual

Life course study has been defined as the joint statistical analysis of three dimensions of longitudinal data: the *timing* of events (when events happen), their *sequencing* (in which order events happen), and their *quantum* (how many events happen) [9]. Two families of methods have been developed to address the complexity of such trajectories and effectively cope with them.

Event-history analysis (EHA) focuses on the time to event, estimating the time at which a transition between two states happens. In epidemiology, EHA is widely used to parametrically estimate the hazards of disease or death, while examples of EHA application in other fields include parenthood, divorce, end of unemployment. If the focus is on an irreversible event, i.e. only one change of state is possible, this family of methods is referred to as survival analysis [10]. When events can repeat, and multiple changes of state are possible, the joint examination of times and changes of states is more challenging. Multistate models allow estimating transition probabilities from each state to multiple destination states, but in their basic form they assume that such transitions rates only depend on the present state, regardless of past behaviour [11]. As opposed to time-to-event estimation, an holistic approach considering series of events as conceptual units has been put forward under the name of SSA [12]. This non-parametric data-mining technique allows describing and summarizing the timing, sequencing, and quantum of life course events gazing the whole picture. While complex life courses have been analyzed via SSA in demographic research for over two decades [1,2], this method has only been recently "discovered" by health researchers, as availability of EHR has grown, with a few applications on diagnostic-therapeutic pathways and pharmacovigilance [13–16].

2.1 SSA for Categorical Discrete Longitudinal Data

Say that for each discretized time unit t (day, time, month, year), each individual i can be associated with a variable S_{it} indicating their life course status at time t. S_{it} can be described by a character denoting one particular state. The state space, which encompasses all possible states an individual can take in each time period, is commonly referred to as *alphabet*. For instance, if we consider the marital status of the respondent, the possible states could be: S for single, C for

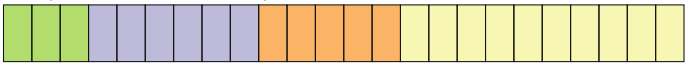
cohabiting, and M for married. The individual's ordered sequence of states, also referred to as *trajectory*, can be written as a string (e.g. SSCCCMM), or as the ordered pairs of successive distinct states S_i and corresponding durations $t_{s,i}$, i.e. (S,2),(C,3),(M,2). After coding the life course in the form of a sequence, the next step of SSA focuses on the identification of common patterns of trajectories. Data reduction techniques such as cluster analysis or multidimensional scaling can assist in defining clusters of life courses by taking as input a SSA-based *dissimilarity matrix*, i.e. a matrix of pairwise dissimilarities between individuals. Alternatively, when the interest is on quantifying how far an individual is from an "ideal" behaviour, a distance from a reference trajectory can be computed. Please refer to [17] for a complete description of sequence distances and clustering.

3 Application to Childhood Vaccination

The first months of a newborn's life are a critical period for their development, including in terms of their immune system [18]. Vaccination is the most effective strategy to boost immunity against infectious disease, protecting infants from life-threatening infections during their first months of life. Optimal timing of vaccination is critical to maximize protection, providing specific antibodies during the time window in infancy with heightened susceptibility to severe disease. Although immunization policies promote a uniform healthcare intervention—similar vaccine doses are administered at the same time to all people—some refuse vaccines or delay uptake [19]. Beyond the individual protection, public health measures must ensure the attainment of a target vaccination coverage to protect through herd immunity the small number of children who cannot receive the vaccine. As such, accurately monitoring the parental decisions regarding their child's vaccine uptake and timing is of foremost importance. Increasingly prevalent EHR allow to efficiently gather vaccination and demographic data, and to profile the population groups that are under-covered by the national programmes in order to subsequently inform evidence-based interventions. In particular, an analysis of individual-level vaccination trajectories can add to the field by describing the diversified pathways in access and timeliness.

3.1 Vaccination Trajectories

Similarly to life course trajectories, we propose to analyze vaccination trajectories making use of SSA. Assuming of observing each individual for $T = 24$ time points (e.g. for 24 months), a sequence of states will look like



where green = 0 doses, violet = 1 dose, orange = 2 doses, yellow = 3 doses. While the invariable order of states (i.e. dose 2 can only happen after dose 1) makes vaccination sequences somewhat simpler with respect to other life course trajectories, a degree of variability across individuals is expected in terms of timing, e.g. when uptake of each dose happens, as well as in terms of quantum.

Some individuals may interrupt the cycle after taking up one or two doses, while other individuals may avoid vaccination completely. Finally, it is realistic to assume that timings of different doses are related to each other, not only because a delay in the first dose implies subsequent delays, but also because individuals delaying the first dose may represent a more skeptical group.

3.2 Data and Findings

We accessed individual-level EHR for hexavalent vaccination in the Lombardy region, where the public healthcare system covers over 10 million people. Hexavalent vaccine (protecting from diphtheria, tetanus, pertussis, poliomyelitis, haemophilus B, and hepatitis B) is routinely administered as a cycle of three doses during the 3rd, 5th and 11th months from birth [20]. For all children born between 2006 and 2019, a total of 810,570 individuals, we can build month-specific trajectories of vaccination up to the second birthday.

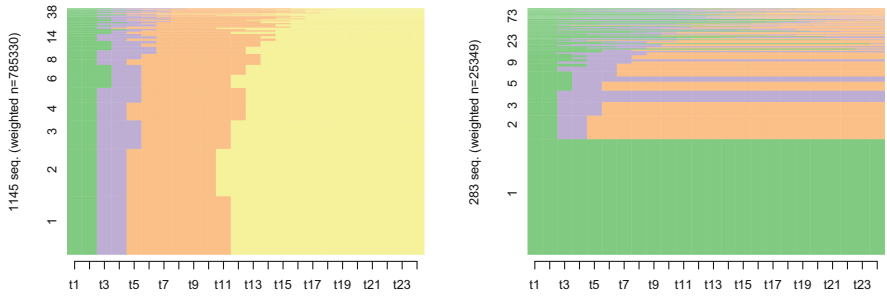


Fig. 1. Sequence index plot for trajectories of hexavalent vaccination. Colors represent the number of doses received. Green=0, violet=1, orange=2, yellow=3 (Color figure online)

The sequence plots in Fig. 1 summarize the individual longitudinal vaccination paths, while the width of each bar represents the prevalence of each sequence in our population. Most common trajectories are depicted bottom to top. The left panel presents trajectories for the individuals who received the complete immunization cycle by their second birthday, 96.87% of our cohort. The most common vaccination trajectory corresponds to the optimal schedule promoted by public health officers: 22% of our cohorts precisely adhered to it. The other sequences in this panel, mostly similar to the first one, show delays of few months across doses. The right panel summarizes trajectories of the remaining 3.13% of our sample. Zero-dose children, in green, lack documentation of receipt of any dose of hexavalent, and they represent 1.4% of our cohort, whereas another 1.73% of children interrupted the cycle after taking one or two doses; most children who didn't receive the third (bar ending in orange) also had delayed administration of the first and/or second dose.

Finally, we provide some synthetic population-level information. Figure 2 shows, on the left panel, the empirical time-specific transition rates: the highest rates for the 0→1 transition, i.e. uptake of the first dose, are observed at months 3 and 4, with rates sharply declining right after. Transitions 1→2 and 2→3 are more dispersed, the first showing high prevalence between the 5th and the 9th month, and the latter between the 12th and the 15th month, both featuring right tails indicating delays. The state distribution plot on the right panel, showing the prevalence of each state at each time point, signals the considerable number of individuals who haven't received the third dose at months 13–20, on top of "interrupters" and zero-dose children described above.

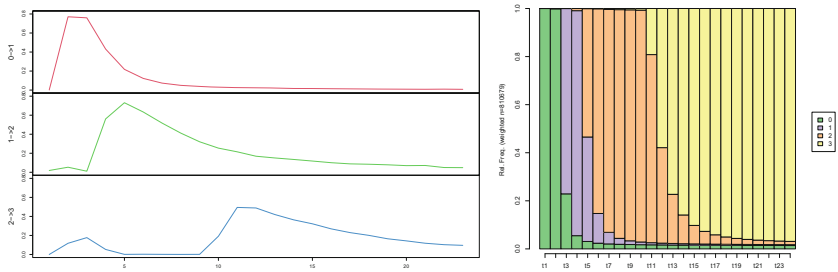


Fig. 2. Left panel: time-specific rates of hexavalent vaccination. Right panel: state distribution plot

3.3 Discussion and Future Steps

Different health institutions monitor the hexavalent vaccine coverage. The WHO Global Health Observatory reports the percentage of one-year-olds who received three doses for each antigen: latest data for the European region indicate a 94% coverage for diphtheria, tetanus and pertussis (DTP) and polio in 2022 [5]. Epicentro, the Italian surveillance system, reports vaccination coverage at the second birthday instead [6]. However, no previous work has leveraged detailed individual-level data from EHR to identify the standard trajectories, and the departures from these standards, disentangling the behaviour in terms of timeliness and going beyond the dichotomous measures of complete vs incomplete, on time vs delayed. Showing patterns of uptake, interruption, or complete refusal, we emphasize the risk of introducing biases when not taking timing into account.

Nevertheless, the role played by exogenous factors such as parental beliefs and sociodemographics on timeliness should be further investigated: they have been shown to impact vaccine refusal, but evidence is limited when it comes to the risk of delaying or interrupting the primary vaccination cycle. A comparison of trajectory evolution across birth cohorts could also uncover secular trends.

4 Closing Remarks

We have shown how an holistic approach can assist in the analysis of vaccination uptake, where trajectories are characterized by pre-determined sequencing, i.e.

no backward transitions, and the relevant research questions focus on timing and quantum. Beyond the specific example of hexavalent vaccine, this framework can be applied to other multi-dose vaccines as well as increasing-only processes. In addition to the graphical summary, SSA can measure differences between trajectories and identify homogeneous groups of typical trajectories through clustering. In this work, for the sake of space, we have not touched on the choice of dissimilarity metrics nor clustering methods. Furthermore, clusters obtained from the analysis of sequences of one phenomenon, e.g. timeliness of primary vaccination cycle, could also be employed to infer future health behaviours, such as the uptake of other vaccines.

References

1. Billari, F.C.: Sequence analysis in demographic research. Can. Stud. Population **28**(2), 439–458 (2001)
2. Jalovaara, M., Fasang, A.E.: Family life courses, gender, and mid-life earnings. Eur. Sociol. Rev. **36**(2), 159–178 (2020)
3. Zhang, C.X., et al.: Ethnic inequities in routine childhood vaccinations in England 2006–2021: an observational cohort study using electronic health records. Eclinicalmedicine **65** (2023)
4. Truong-Vu, K.P.: On-time, late, or never: incorporating intersectionality to predict age-specific probabilities of initiating the HPV vaccine series. Soc. Sci. Med. **287**, 114227 (2021)
5. https://immunizationdata.who.int/listing.html?topic=coverage&location=
6. https://www.epicentro.iss.it/vaccini/datiIta
7. Moore, H.C., et al.: Assessment of on-time vaccination coverage in population subgroups: a record linkage cohort study. Vaccine **36**(28), 4062–4069 (2018)
8. Suffel, A., et al.: Timeliness of childhood vaccination in England: a population-based cohort study. Vaccine **41**(39), 5775–5781 (2023)
9. Billari, F.C., Fürnkranz, J., Prskawetz, A.: Timing, sequencing, and quantum of life course events: a machine learning approach. Eur. J. Population **22**, 37–65 (2006)
10. Clark, T.G., et al.: Survival analysis part I: basic concepts and first analyses. Br. J. Cancer **89**(2), 232–238 (2003)
11. Cadarso-Suárez, C., et al.: Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data. Stat. Model. **10**(3), 291–314 (2010)
12. Liao, T., et al.: Sequence analysis: its past, present, and future. Soc. Sci. Res. **107**, 102772 (2022)
13. Vanoli, J., et al.: Use of State sequence analysis in pharmacoepidemiology: a tutorial. Int. J. Environ. Res. Public Health **18**(24), 13398 (2021)
14. Baudrier, C., et al.: Identifying homogeneous healthcare use profiles and treatment sequences by combining sequence pattern mining with care trajectory clustering in kidney cancer patients on oral anticancer drugs: a case study. Health Inf. J. **28**(2), 14604582221101526 (2022)
15. Flothow, A., et al.: Analytical methods for identifying sequences of utilization in health data: a scoping review. BMC Med. Res. Methodol. **23**(1), 212 (2023)
16. Savaré, L., et al.: Capturing the variety of clinical pathways in patients with schizophrenic disorders through state sequences analysis. BMC Med. Res. Methodol. **23**(1), 174 (2023)

17. Abbott, A., Forrest, J.: Optimal matching methods for historical sequences. J. Interdisc. Hist. **16**(3), 471–494 (1986)
18. Olin, A., et al.: Stereotypic immune system development in newborn children. Cell **174**(5), 1277–1292 (2018)
19. Chiavenna, C., et al.: Personal risk or societal benefit? Investigating adults' support for COVID-19 childhood vaccination. Vaccine **41**(25), 3683–3687 (2023)
20. https://www.wikivaccini.regione.lombardia.it/wps/portal/site/wikivaccini/DettaglioRedazionale/le-vaccinazioni-quali/calendario-vaccinale



A Flexible Parametric Approach to Synthetic Patients Generation in Clinical Trials

Marta Cipriani[1,2(✉)], Lorenzo Di Rocco[1], and Marco Alfò[1]

[1] Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
[2] Mathematical Institute, Leiden University, Leiden, Netherlands
marta.cipriani@uniroma1.it

Abstract. Enhancing research reproducibility and data accessibility is essential in scientific research. However, ensuring data privacy while achieving these goals is a challenging task, especially in the medical field where sensitive data are often commonplace. One solution is to use synthetic data that mimic real-world datasets. This method has the potential to streamline therapy evaluation and facilitate quicker access to innovative treatments.

We propose an approach that uses sequential conditional regression and flexible parametric survival models to accurately replicate covariate patterns and survival times. The approach is easily implementable through an R function. We also provide an application to an onco-hematological trial. The results show the potentialities of the proposed method in mirroring original distributions and survival outcomes.

Keywords: synthetic · flexible parametric survival model · simulation

1 Introduction

Artificial Intelligence (AI) is transforming numerous medical fields, such as image analysis, multi-omics data integration, drug discovery and precision medicine. Some notable applications have already been explored in these fields [1,2]. One promising yet relatively unexplored frontier is the use of synthetic data in clinical trial design.

Initially introduced to provide publicly accessible alternatives to original datasets, synthetic data holds significant potential in clinical trial applications. It can create a virtual patient group that closely mimics real-world datasets while protecting individual patient privacy. This is achieved by replacing observed values with values sampled from appropriate probability distributions to preserve the statistical characteristics of the original data. By adopting this approach, the barrier to open scientific discourse may be broken down, and limitations associated with common anonymisation practices are overcome. Notably, ensuring the efficacy of anonymisation has become increasingly challenging in the face of ongoing technological advancements [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 419–425, 2025.
https://doi.org/10.1007/978-3-031-64431-3_69

We propose methods to emulate observed survival data and present a user-friendly R function, called `fleSSy`[1] (FLExible Survival SYnthetic), to implement them. Additionally, we discuss the application of these methods to a cohort of patients affected by acute myeloid leukemia (AML) [4].

2 Methods

In the following sections, we will describe the steps that `fleSSy` uses to generate synthetic data.

2.1 Fitting the Survival Model on Observed Covariates

The individual survival times are modelled by the flexible parametric proportional-hazards model proposed by Royston and Parmar [5]. The approach they proposed is to model the logarithm of the baseline cumulative hazard function as a "natural" cubic spline function of log time. Generically speaking, the class of such models is based on a transformation of the survival function $S(t)$ by a link function $g(\cdot)$. For the proportional-hazards spline model with fixed covariate vector $\mathbf{z}$ we have

$$g[S(t;\mathbf{z})] = \log[-\log S(t;\mathbf{z})] = \log H(t;\mathbf{z}) = \log H_0(t) + \beta^T\mathbf{z} = s(x;\gamma) + \beta^T\mathbf{z} \tag{1}$$

where $H(t;\mathbf{z})$ is the cumulative hazard function, $H_0(t)$ is the baseline cumulative hazard function, and $s(x;\gamma)$ is a natural cubic spline. The latter is defined as a cubic spline constrained to be linear beyond boundary knots $k_{\min}, k_{\max}$. In addition, m distinct internal knots $k_1 < \cdots < k_m$ are specified, and a natural cubic spline may be written as

$$s(x;\gamma) = \gamma_0 + \gamma_1 x + \gamma_2 \upsilon_1(x) + \cdots + \gamma_{m+1}\upsilon_m(x) \tag{2}$$

where the jth basis function $\upsilon_j(\cdot)$ is defined for $j = 1, \ldots, m$ as

$$\upsilon_j(x) = (x - k_j)_+^3 - \lambda_j(x - k_{\min})_+^3 - (1 - \lambda_j)(x - k_{\max})_+^3 \tag{3}$$

where $\lambda_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}$ and $(x - a)_+ = \max(0, x - a)$.

2.2 Generation of Synthetic Covariates Distributions

To reconstruct the joint distribution of the covariates that are used within the survival model, we used a fully conditional specification (FCS) approach [6]. This method helps us avoid explicitly specifying the joint distribution and it rather proposes to build the full joint distribution using conditional distributions.

We used an approach similar to the one adopted in the multiple imputation by chained equations (MICE) algorithm [7]. The following method is implemented

[1] The source code is available at https://github.com/ldirocco/fleSSy.

in `fleSSy` using the pre-existing features of the `syn` function in the `synthpop` R library [8].

Let X be a set of p covariates, and let $X^{(j)}$ $(j = 1, \ldots, p)$ denote the jth covariate in X, with $X^{(-j)}$ denoting all covariates in X except $X^{(j)}$. We used the distribution of $X^{(j)}|X^{(-j)}$ to generate the set of synthetic values for the jth covariate.

The form of the regression model depends on the variable type. We used a logistic regression model for binary variables, a polytomous logistic regression for factors with more than two levels, and an ordered polytomous logistic regression for ordered factors with more than two levels. For continuous variables, we employed a linear regression model with an inverse normal-rank-based transformation of the covariate of interest to approximate continuous covariates with non-normal distributions.

To generate synthetic data which closely mimics the original data, we required two essential information in our initial design matrix: a binary variable indicating the dropout status and the time of entry into the study for each individual. The user can choose which additional covariates to synthesize.

2.3 Generating Simulated Survival Times

After generating the complete set of covariates, we simulate survival times using the survival model we have previously fitted to the original data. We consider the date of the end of the study as our right censoring time.

The subsequent step is to code the survival status by combining dropout, right-censoring, and death.

It is important to note that we synthesize each individual's covariate pattern from a distributional range, rather than a specific real-world individual. Therefore, we do not model survival times directly from any real person. Consequently, there is no connection between a synthetic individual and any individual in the real-world dataset.

2.4 Comparison Between the Original and Synthetic Distributions

We used rigorous statistical methods to examine how well the synthetic cohort matched the original one. To achieve this, we combined the original and synthetic records to measure how accurately the data values predict the source of the records, using the propensity score - the predicted probability that a record originates from the synthetic data.

Following the recommendation of Snoke *et al.* [9], the most commonly suggested utility measure in this context is the propensity score mean squared error (pMSE) or its standardized ratio. Drawing upon the works of Raab *et al.* [10], we propose a practical threshold for assessing utility: if all standardized pMSE ratios fall below 10, and preferably below 3, further adjustments may be deemed unnecessary, indicating the satisfactory utility of the synthetic data.

Another approach to utility measures involves grouping the original and synthetic data by constructing tables based on their values and computing measures of difference between these tables. We used the Welch two-sample t-test, the Mann-Whitney test and the two-sample test for equality of proportions to evaluate adherence across both continuous and categorical variables. Moreover, we evaluated the differences in the survival outcomes of the virtual and actual cohorts using the Log-rank test.

Recognizing that a high degree of statistical similarity may not necessarily imply true similarity, especially when considering the intricate interplay between variables, and appreciating the critical importance of these interrelationships, we conducted a thorough stratified survival analysis. This enabled us to gain further insights into the influence of factors on distinct patient subgroups.

3 Case Study

The feasibility of the proposed methods of synthetic patient generation is shown through an application to data collected in a phase II experimental trial, the GIMEMA AML1310, carried out by Fondazione GIMEMA Franco Mandelli Onlus.

3.1 Available Data

The GIMEMA AML1310 study enrolled 445 patients across 55 Italian haematology institutions. It was a trial for newly diagnosed AML patients which included a "3+7"-like induction and a risk-based MRD-directed post-remission transplant allocation.

The covariates synthesised included age, sex, white blood cells, ELN2017 risk category, dropout rate, and time since the start of the study.

3.2 Results

The synthetic data closely resembles the covariate patterns found in the original data, ensuring that survival predictions conditional on distinct covariate profiles are appropriately reflected.

Table 1 provides as summary of the features of the synthetic cohort and compares them to those of the original population. Notably, the clinic-biological characteristics of the two cohorts did not differ significantly, at least based on the central measures. However, further insights may be gained by examining t-distribution-free test statistics on the cumulative distribution function (CDF).

Remarkably, for every synthesized variable, the associated standardized pMSE (S_pMSE) falls well below the suggested threshold.

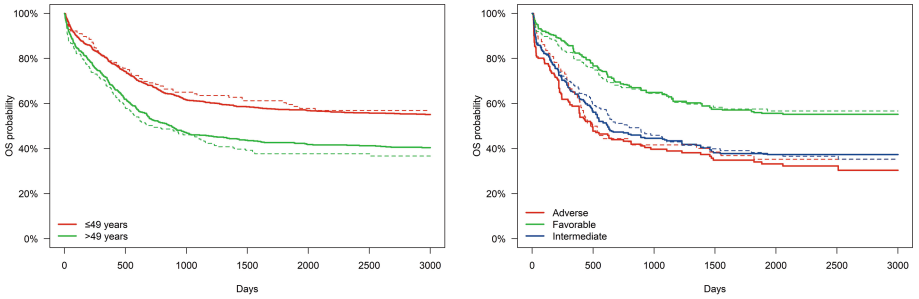
Moreover, the effectiveness of these methods in replicating survival patterns found in the original data is demonstrated by comparing various survival estimates for the original and synthetic data. The marginal OS curves demonstrated striking similarity, with a p-value of 0.6 when comparing original and synthesised survival curves through a Log-rank test.

Table 1. Comparison between the original and synthetic AML1310 cohorts in terms of demographic, clinic-biologic characteristics, and survival response.

Characteristic	Original	Synthetic	p-value[a]	S_pMSE[b]
Survival status, n (%)				0.76
Alive/Censored	217 (48.8%)	211 (48.3%)	0.74	
Dead	228 (51.2%)	226 (51.7%)	0.95	
Age, median (range)	49 (18, 61)	50 (18, 61)	0.91	0.94
Sex, n (1%)				1.09
Male	232 (52.1%)	243 (54.6%)	0.50	
Female	213 (47.9%)	202 (45.4%)	0.50	
WBC, median (range)	14 (0, 341)	13 (0, 341)	0.73	0.17
ELN2017, n (%)				0.24
Adverse	80 (18.0%)	79 (17.8%)	>0.99	
Intermediate	179 (40.2%)	173 (38.9%)	0.73	
Favorable	186 (41.8%)	193 (43.4%)	0.68	
Dropout, n (%)	32 (7.2%)	36 (8.1%)	0.71	1.50
Time from start, median (range)	783 (0, 1230)	810 (0, 1230)	0.66	1.18

[a] Welch Two-Sample t-test; Mann-Whitney test; Two-sample test for equality of proportions

[b] Standardised propensity score Mean Square Error



(a) OS stratified by age class

(b) OS stratified by risk categories

Fig. 1. Survival estimates of the original and synthetic cohorts. Dashed lines refer to the original cohort and continuous lines to the synthetic patients.

To further underscore the agreement in survival estimates, OS curves were stratified by age class (as shown in Fig. 1.a.) and risk categories (as shown in Fig. 1.b.). In both cases, the survival estimates for the synthetic (continuous lines) and original (dashed lines) cohorts are very similar.

4 Conclusions

This work shows the huge potential derived from creating a virtual group of patients using real patient data, with a specific focus on an AML population. The resulting synthetic cohort may be used as a control group for testing new treatments through the so-called *in-silico* trials [11] or, even more importantly, can be distributed without privacy-related issues, therefore stimulating the debate on replicability. Indeed, while randomized controlled trials are still considered the gold standard, there is a growing recognition of the need for alternative approaches that can help when timely decisions are needed.

The benefits of using synthetic patient data include minimizing patient exposure to inactive treatments, faster trial completion, cost optimization, and enhanced privacy protection. Furthermore, synthetic patient generation also addresses limitations such as small sample sizes and covariate imbalances, which may help improve efficiency in treatment effect estimates and predictive models.

However, there may be regulatory hurdles and legal uncertainties that could impede the widespread adoption of these techniques. At the same time, non-standard statistical issues, mainly linked to the representativeness of the sample at hand, may arise.

Future developments of this work involve incorporating the `fleSSy` function into a comprehensive Python library for synthetic survival data generation.

References

1. Grewal, J.K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., et al.: Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. JAMA Netw. Open **2**(4), e192597 (2019). https://doi.org/10.1001/jamanetworkopen.2019.2597
2. Khanani, S.: Editorial comment: artificial intelligence in mammography-our new reality. Am. J. Roentgenol. **219**(3), 381 (2022). https://doi.org/10.2214/AJR.22.27345
3. Rocher, L., Hendrickx, J.M., de Montjoye, Y.A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. **10**(1), 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3
4. Venditti, A., Piciocchi, A., Candoni, A., Melillo, M., et al.: GIMEMA AML1310 trial of risk-adapted, MRD-directed therapy for young adults with newly diagnosed acute myeloid leukemia. Blood J. Am. Soc. Hematol. **134**(12), 935–945 (2019). https://doi.org/10.1182/blood.2018886960
5. Royston, P., Parmar, M.K.B.: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat. Med. **21**, 2175–2197 (2002). https://doi.org/10.1002/sim.1203
6. van Buuren, S., Brands, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B.: Fully conditional specification in multivariate imputation. J. Stat. Comput. Simul. **76**(12), 1049–1064 (2006). https://doi.org/10.1080/10629360600810434
7. van Buuren, S.: Flexible Imputation of Missing Data, Second Edition. Chapman & Hall/CRC Interdisciplinary Statistics (2012)

8. Nowok, B., Raab, G.M., Dibben, C.: synthpop: bespoke creation of synthetic data in R. J. Stat. Softw. **74**(11) (2016). https://doi.org/10.18637/jss.v074.i11
9. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. J. Roy. Stat. Soc. Ser. A Stat. Soc. **181**(3), 663–88 (2018). https://doi.org/10.1111/rssa.12358
10. Raab, G.M., Nowok, B., Dibben, C.:Assessing,visualizing and improving the utility of synthetic data. arXiv preprint arXiv:2109.12717 (2021)
11. Pappalardo, F., Russo, G., Tshinanu, F.M., Viceconti, M.: In silico clinical trials: concepts and early adoptions. Brief. Bioinform. **20**(5), 1699–708 (2019). https://doi.org/10.1093/bib/bby043

A Multi-Domain Model with Partial Differential Regularization: An Application to Neuroimaging Data

Aldo Clemente[1(✉)], Eleonora Arnone[2], and Laura M. Sangalli[1]

[1] MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano20133, Italy
aldo.clemente@polimi.it, laura.sangalli@polimi.it

[2] Department of Management, Università degli Studi di Torino, Corso Unione Sovietica 218 bis, Torino 10134, Italy
eleonora.arnone@unito.it

Abstract. In the framework of physics-informed statistical models, this work proposes a multi-domain spatial regression method with a regularization term involving the Laplace-Beltrami operator, specifically designed for data observed over surface domains. To illustrate its application, we employ the proposed method on high-dimensional resting-state fMRI signals from various subjects, with the precuneus designated as the Region of Interest for computing the Functional Connectivity maps. Here, we treat the Functional Connectivity Map of each subject as the response variable, with available data on cortical thickness serving as the regressor.

Keywords: multi-domain · finite element · neuroimaging

1 Introduction and Methodology

Neuroimaging is a multidisciplinary field that involves various techniques to visualize and study the structure, function, and connectivity of the brain. Typically, the analysis of functional neuroimaging signals occurs on a voxel-wise basis, within the framework known as mass univariate analysis. Nonetheless, the inherent limitations in power within neuroimaging studies have prompted neuroscientists to devise novel approaches to enhance this framework. Various strategies have been devised to incorporate spatial variation into statistical inference or address deviations from the assumed data distribution. See [1] for an overview on statistical methods in neuroimaging. This work contributes to the literature on Spatial and Functional Data Analysis for neuroimaging data. Examples of FDA methods for neuroimaging data can be found in the literature, such as in [2–5].

In the context of Functional Neuroimaging group studies, this work proposes a multi-domain spatial regression method with Partial Differential Equation (PDE) regularization for data observed over surface domains.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 426–430, 2025.
https://doi.org/10.1007/978-3-031-64431-3_70

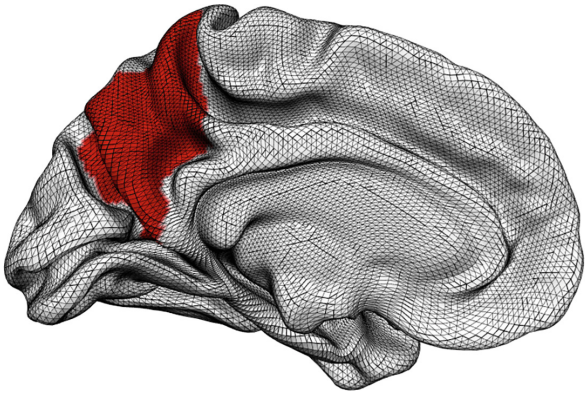


Fig. 1. Template cortical surface mesh used in the application. The precuneus (ROI) is highlighted in red. (Color figure online)

Our objective is to analyze fMRI data originating from the cerebral cortex, which is the intricately folded thin layer where the majority of neural activity occurs. We can describe this thin layer as a two-dimensional surface embedded in a three-dimensional space, characterized by a 2D geodesic distance, rather than the 3D Euclidean distance within the volume. Figure 1 shows the triangulation of the template cortical surface. The PDE regularization within the proposed method accounts, indeed, for the complex shape of the cortical surface where data are observed. In such respect, the proposed method can be seen as an extension of the class of physics-informed statistical methods reviewed by Sangalli [6].

2 Multi-Domain Model with Partial Differential Regularization

Denote with Ω_i the surface domain of the i-th patient embedded in the three-dimensional space. For each patient $i = 1, \ldots, m$, we observe a quantity of interest y_{ij} at location $\mathbf{p}_{ij}$, along with a set of group-effect covariates $x_{ij}^\top \in \mathbb{R}^q$ and a set of patient-specific covariates $v_{ij}^\top \in \mathbb{R}^p$. We model y_{ij} as

$$y_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta} + v_{ij}^\top \boldsymbol{\alpha}_i + f_i(\mathbf{p}_{ij}) + \varepsilon_{ij},$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is the unknown vector of group-effect coefficients, $\boldsymbol{\alpha}_i \in \mathbb{R}^p$ is the unknown vectors of patient-specific coefficients for the i-th patient, $f_i : \Omega_i \to \mathbb{R}$ is the unknown spatial field for the i-th patient, and ε_{ij} are uncorrelated errors with zero mean and finite variance. We assume that f_i, for all $i = 1, \cdots, m$, belongs to the space of integrable functions up to the second derivatives, i.e., the Sobolev space $H^2(\Omega_i)$, see, for instance, [7] for an introduction to Sobolev spaces. Moreover, we assume that $\sum_{i=1}^m \boldsymbol{\alpha}_i = \mathbf{0}$. To estimate $\boldsymbol{\beta}, \boldsymbol{\alpha}_i$ and f_i for $i = 1, \ldots, m$, we solve the following constrained minimization problem:

$$\begin{cases} \underset{\boldsymbol{\beta},\boldsymbol{\alpha}_i,f_i}{\arg\min} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \left(y_{ij} - \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} - v_{ij}^{\top}\boldsymbol{\alpha}_i - f_i(\mathbf{p}_{ij}) \right)^2 + \lambda \int_{\Omega_i} \left(\Delta_{\Omega_i} f_i \right)^2 d\Omega_i \right) \\ \sum_{i=1}^{m} \boldsymbol{\alpha}_i = \mathbf{0}, \end{cases} \tag{1}$$

where Δ_{Ω_i} is the Laplace-Beltrami operator. The first term of the first equation in (1) represents the sum of squared errors while, the second term, involving the Laplace-Beltrami operator, serves as a regularization term, penalizing the roughness of the estimates. The minimization of problem (1) can be approached either directly or through an iterative method. Specifically, the iterative approach is well-suited for addressing large-scale problems, such as those encountered in neuroimaging applications. A possible future development of this work would involve investigating the embedding of the patient-specific structural information of the brain. In this framework we could leverage, for instance, the patient-specific Diffusion Tensor Image (DTI) in the minimization of problem (1). See Zhu et al. [8] for a review on methods that fuse fMRI and DTI modalities.

3 Analysis on Resting-State fMRI Data

The resting-state fMRI data considered in this work originate from the Human Connectome Project Consortium [9]. For both the alignment of the patient-specific cortical surface to the template cortical surface and the pipeline for data preprocessing, refer to [10]. The data considered in this application have previously been examined within the framework of Functional Principal Component Analysis (fPCA) by Lila et al. [2]. A common method for analyzing resting-state fMRI data involves the selection of a Region of Interest (ROI) on the cortical surface, followed by computing an average time series within each subject's ROI. Finally, the correlation between the time series over the whole cortical surface and the ROI time series is computed. The resting-state functional connectivity map is obtained by applying a Fisher's r-to-z transformation to the correlation coefficients. Therefore, for each patient, the processed data include a resting-state Functional Connectivity map over the left hemisphere of the cortical surface and measurements of cortical thickness. In our analysis, we consider the resting-state Functional Connectivity Map as the response variable and we use the cortical thickness as both group-effect and patient-specific effect. The data locations correspond to the mesh nodes. Figure 1 shows the template surface mesh, composed of almost 32k nodes, used in this application, and the ROI is highlighted in red.

The proposed method estimates smooth patient-specific functions defined over the cortical surface of the brain. Figure 2 shows the patient-specific estimated spatial fields displaying different patterns. A 95% confidence interval, $IC_{0.95} = [0.0055 \pm 0.00069]$, has been constructed using a bootstrap approach for β.

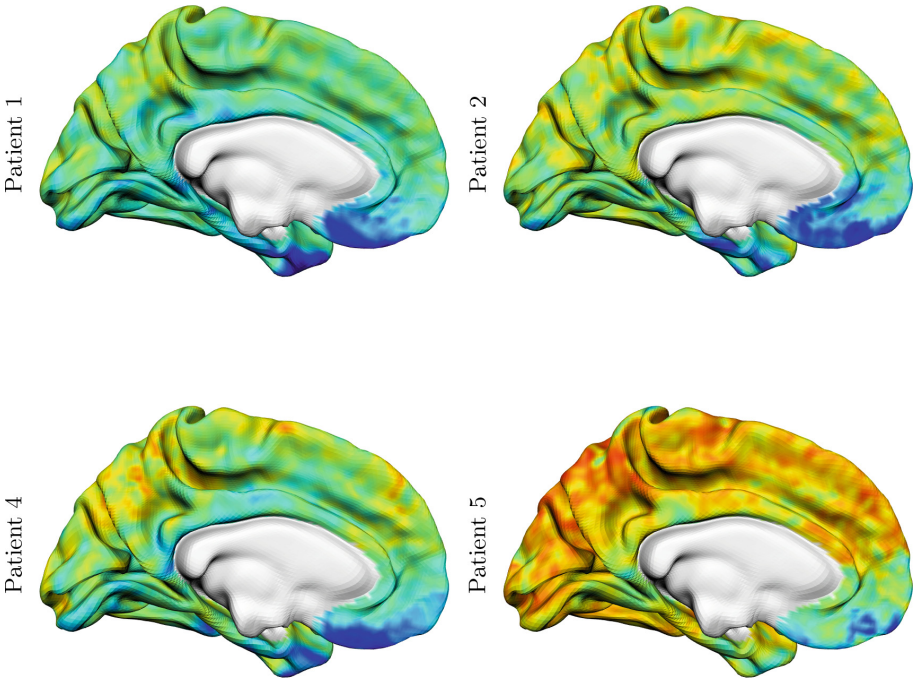


Fig. 2. Spatial field estimates $\hat{f}_i$, for patients $i = 1, 2, 4, 5$.

Acknowledgements. Aldo Clemente acknowledges funding by by European Union - NextGenerationEU. We also acknowledge the MUR research project Dipartimento di Eccellenza 2023–2027, Dipartimento di Matematica, Politenico di Milano.

References

1. Ombao, H., Lindquist, M., Thompson, W., Aston, J.: Handbook of Neuroimaging Data Analysis. CRC Press (2016)
2. Lila, E., Aston, J.A., Sangalli, L.M.: Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. Ann. Appl. Stat. **10**(4), 1854–1879 (2016)
3. Lila, E., Aston, J.A.: Functional random effects modeling of brain shape and connectivity. Ann. Appl. Stat. **16**(4), 2122–2144 (2022)
4. Arnone, E., Negri, L., Panzica, F., Sangalli, L.M.: Analyzing data in complicated 3D domains: smoothing, semiparametric regression, and functional principal component analysis. Biometrics **79**(4), 3510–3521 (2023)
5. Clementi, L., Arnone, E., Santambrogio, M.D., Franceschetti, S., Panzica, F., Sangalli, L.M.: Anatomically compliant modes of variations: new tools for brain connectivity. PLoS ONE **18**(11), e0292450 (2023)
6. Sangalli, L.M.: Spatial regression with partial differential equation regularisation. Int. Stat. Rev. **89**, 505–531 (2021)

7. Brezis, H., Brézis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations, vol. 2. Springer (2011)
8. Zhu, D., et al.: Fusing DTI and fMRI data: a survey of methods and applications. Neuroimage **102**, 184–191 (2014)
9. Van Essen, D.C., et al.: The human connectome project: a data acquisition perspective. Neuroimage **62**(4), 2222–2231 (2012)
10. Glasser, M.F., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013)



Standardization Process of Budget Administrative Data, Having Statistic Characteristics and Usefulness for Public Finance Control

Daniela Collesi(✉)

Italian Ministry of Economy and Finance, Rome, Italy
daniela.collesi@mef.gov.it

Abstract. The accounting reform introduced in Italy by Law 196 of 2009 has, among its objectives, the harmonization of accounting. It defines a series of rules that make possible to produce data ready to be used both for the administrative purposes of government and for production of public finance statistics. To achieve this goal the decision was to adopt, in the administrative area of public budgets - for all phases of preparation, management and reporting - the classifications of official statistics, according to the national accounts' rules. The introduction of the new classifications was significant and required a four-year experimental process for the Italian local authorities. The expenditure classification by Missions and programs, associated with the international COFOG[1] classification, is the main objective of this document also taking into account that a glossary was prepared that defines in detail the content of each mission, program and the corresponding COFOG.

The budget data for the five-year period 2016–2020 show interesting results that confirm both the importance of the direct adoption of the standard classifications of the national accounts into the budget and the ability of the local authorities to improve over time, as evidenced by the increasing quality of the data in the last two years, compared to the 2016 data, the first year the accounting reform was applied to all local authorities.

Keywords: adoption of statistical classifications · cross-institutional comparability · expenditure budget data · new budget rules

1 Harmonisation Reform Introduced by Law 196/2009

The harmonisation of public budgets has been a recurring theme over time, linked to the need to have comparable data between public administrations for the purposes of consolidating the general government account, constructing/implementing a homogeneous information base in the definition of budget figures and in the way they are represented.[1]

[1] COFOG (Classification Of Function Of Government) is the international classification of public expenditure by function/purpose; it is part of the UN statistical classification by purpose, and it foresees 3 levels of classifications, the first concerns 10 divisions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 431–437, 2025.
https://doi.org/10.1007/978-3-031-64431-3_71

The need for harmonisation was a common element for all the institutional stakeholders who use public accounts in the hearings held in Parliament on the examination of the proposed law on accounting and public finance; it is therefore acknowledged that the provisions contained in Article 2 of Law No. 196/2009 derive from the prior request and approval by the main stakeholders of public budgets and their synthesis in statistical terms. The National Institute of Statistics fully agreed with the harmonisation process extended to the entire general government sector, providing for the adoption of common budget schemes split by missions and programmes for the expenditure, since it allows to represent the framework of public policies implemented by general government through the direct link to the economic and functional classifications of the ESA, thus enabling quantitative data to be consolidated according to a nomenclature that ensures homogeneous contents.

2 The Classification Changes

With the support of a Committee for the Accounting Principles of Public Administrations, rules for implementing the delegation given by the public finance and accounting law (Law 196/2009) and to coordinate the harmonisation system were set up. The implementation of the discipline on accounting harmonisation provided that the public administrations should adopt a single economic classification in line with the operations of the general government account defined by the ESA (National accounts) and that the classification by Missions and Programmes (purposes of expenditure) should be accompanied by the COFOG classification, a classification of the functions of public administrations. The biggest challenge was the application of these new accounting rules to the territorial administrations, more than 9,000 entities, and within them to the regions, 20 regions and two autonomous provinces, administrations that until then had adopted different accounting rules. A great help in this process came from a constitutional amendment to Article 117, which brought accounting matters and the harmonisation of public budgets back under the exclusive competence of the State as of 2014, whereas previously the regions had been granted autonomy in this field.

To respond to the need to support administrations in their accounting revolution, a four- year experimental process was launched for local authorities - regions, provinces, municipalities, and their subsidiaries bodies - which make up the majority of Italian public bodies. This process has enabled the tools available to them to be refined, including through a subsequent permanent commission that analyzes the main critical issues, proposing common shared solutions aimed at ensuring uniformity of the rules for all the bodies falling within the scope considered. For expenditure, which provides for identification of the purpose through the definition of Missions and programs, associated with the international COFOG classification, a glossary, available in https://www.rgs.mef.gov.it/_Documenti/VERSIONE-i/E-GOVERME1/ARCONET/Glossari/Glossario-delle-missioni-e-dei-programmi-fino-al-2024.pdf, was prepared that defines indetail the content of each program and the corresponding COFOG [1].

3 Use of Uniform Classifications for Administrative Purposes for All Public Administrations

The main choice for the accounting harmonisation has been to use uniform classifications for all administrations, which can be easily and directly linked to the statistical classifications of the national accounts (ESA); therefore, statistical tools have been brought into an administrative context, i.e. into the environment of public resources management. The reason for this choice has to be found in the legislator's requirement to have a direct link to the programmatic objectives of public finance as early as the public budgets, simplifying the procedures for reprocessing budget data to transform them into national accounts data; the budget data had to be more readable according to the transparency principles of the data and of the relevant accounting balances that are fundamental for public finance control and this had to take place as early as the administrative level without the need to process them further by public finance experts.

4 Mission/Programme Classifications and COFOG: Critical Issues Due to the Non-authorisation of COFOG

The budgets of territorial governments, examined in this paper, were redefined on the basis of accounting harmonisation to be approved by political bodies according to the purposes of expenditure, i.e. on the basis of missions and programmes in order to have uniformity with the similar process that had already been defined for the State Budget; the new discipline required that budgets also report the association between missions and programmes with COFOG at the second level, making this classification compulsory for governments but not binding when approving expenditure, as analyzed into Collesi D. (2013–2014) [2]. The absence of the authorisation requirement, i.e. not subject to the approval of the political body, has determined limited attention being paid to this classification aspect than to the missions and programmes that are, instead, an essential requirement of the budget.

To be fully aware of the importance of this aspect therefore requires the involvement of administrations in examining the results that can be deduced from a joint interpretation of all the budgets of the entities belonging to the same group: regions, municipalities, provinces and all the other entities in the sector. In our experience, the way in which the results deriving from the use of the data that the authorities send to the database of public administrations are presented is very important; actually, a proactive approach is needed in their analysis, highlighting the improvements that have been noted over time, recognising the effort that the authorities have made in adopting the new classification structures and pointing out, again in a proactive manner, possible control mechanisms that allow for better quality data.

In our case, an example of this divergent interpretation is related to the data on health care expenditure - a regional responsibility - which appear different depending on the perspective used: whether the missions/programmes or the COFOG functions even if it should be the same into the two perspectives. Both from the overall analysis and from the analysis by region, it is evident that COFOG has not been analysed specifically by the individual entities, while the missions and programmes reveal a "conscious" trend

in the representation of the specific underlying public policies as the programmes are the units by which the budget is approved. The two tables below show the percentage distribution of expenditure according to Missions (the main are showed) and to COFOG Division (an additional division has been added related to expenditure not classified by COFOG). Data are taken from the BDAP (Tables 1 and 2).

Table 1. Total regional expenditure % distribution by Missions

Missions	**2016**	**2017**	**2018**	**2019**	**2020**
01-Institutional, general and management services	11,10%	10,50%	11,00%	11,30%	9,50%
04-Education and right to study	3,20%	3,30%	3,30%	3,40%	3,20%
13-Health protection	***37,50%***	***38,10%***	***40,50%***	***38,80%***	***33,30%***
14-Economic development and competitiveness	1,10%	1,00%	1,30%	1,10%	1,60%
15-Labour policies and vocational training	0,80%	0,80%	0,90%	0,90%	0,90%
16-Agriculture, agri-food policies and fisheries	0,50%	0,60%	0,60%	0,70%	0,60%
17-Energy and diversification of energy sources	0,20%	0,20%	0,20%	0,20%	0,20%
18-Relations with other territorial and local governments	2,80%	2,00%	1,90%	1,90%	1,20%
19-International relations	0,10%	0,10%	0,00%	0,10%	0,10%
20-Funds and provisions	0,10%	0,10%	0,10%	0,00%	0,00%
50-Public debt	2,30%	2,40%	2,30%	2,40%	2,30%
60-Financial advances	3,40%	3,50%	2,70%	2,40%	1,90%
99-Services on behalf of third parties	17,60%	17,70%	14,50%	15,90%	17,20%
Missions remaining	19,40%	19,90%	20,50%	20,90%	21,40%
TOTAL	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

The two tables show how different is the percentage of total expenditure attributable to Health, even if it should be exactly the same; when classified by Missions the expenditure is higher as there is any doubt on its attribution. On the other hand, according to COFOG it is more difficult to the accounters as it is shown by both the lower percentage and the high level of expenditure not classified by COFOG, around 10% and more on yearly basis.

Table 2. Total regional expenditure % distribution by COFOG divisions

COFOG	2016	2017	2018	2019	2020
01-GENERAL PUBLIC SERVICES	32,50%	32,20%	29,70%	31,10%	28,90%
02- DEFENCE	0,00%	0,00%	0,00%	0,00%	0,00%
03-PUBLIC ORDER AND SAFETY	1,10%	1,30%	1,40%	1,40%	1,30%
04-ECONOMIC AFFAIRS	7,70%	7,50%	8,30%	8,40%	9,00%
05-ENVIRONMENTAL PROTECTION	3,60%	4,20%	4,40%	4,50%	4,40%
06-HOUSING AND COMMUNITY AMENITIES	1,80%	1,80%	1,90%	1,90%	1,90%
07-HEALTH	***33,40%***	***34,C%***	***3C,4%***	***34,40%***	***35,20%***
08-RECREATION, CULTURE AND RELIGION	1,10%	1,20%	1,30%	1,40%	1,30%
09-EDUCATION	3,00%	3,10%	3,30%	3,30%	3,10%
10-SOCIAL PROTECTION	2,90%	3,20%	3,60%	3,70%	4,00%
-Not classified by COFOG	12,90%	10,90%	9,70%	10,00%	11,00%
TOTAL	**100%**	**100%**	**100%**	**100%**	**100%**

5 Working Group on Experimentation

A Working Group was set up to follow the experimentation for the whole period and, upon conclusion, it was transformed into the ARCONET Permanent Committee, which continues to follow the developments of harmonised accounting, also with a view to comparability.

The working group has prepared the budget schemes that the authorities have adopted and which, in the light of the comments received during the experimentation, have been updated; given the great novelty of the introduction of COFOG, a glossary[2] has also been prepared to support the authorities by providing the correspondence between the missions/programmes and COFOG at the second level. The public administrations also contributed to the success of this tool by clarifying, during the period, the points that were difficult to interpret in the light of their activities and corresponding expenditure. To support administrations in the process of introducing harmonised accounting, a training course on the novelties of the reform was planned, which started in 2015 and covered the whole country.

6 Main Results

The introduction of accounting harmonisation, in line with the completion of the State budget reform, also provided for the creation of a unitary database to collect all relevant data for the public administration (BDAP) needed for monitoring public finance, managed by the Ministry of the Economy and Finance. In addition to the State budget

data, a very significant part of the BDAP is dedicated to the collection of data from local and regional authorities starting from the final adoption of the accounting harmonisation schemes and rules, i.e. from the 2016 budgets. The acquisition of data from more than 9,000 entities concerned required the creation of a series of checks aimed at ensuring the completeness of the data, compliance with squaring checks, consistency between data from different budget schemes, and the quality of the data transmitted. In order not to block the acquisition of budget data, the process of checking the quality of the data has expanded over the years to become blocking for some particularly important aspects as of 2020, in relation to the fact that the transmission of the data is mandatory and the related database is used by institutional users (Ministry of the Interior, Ministry of the Economy and Finance, Court of Auditors) for the controls required by law, for the allocation of resources deriving from transfers and also for the payment of sanctions. Therefore, there have been phases of successive 'rigidity/severity' in the control of data and this has led to an improvement in their quality. Since the aspect of the purpose of expenditure has so far been used to a limited extent for statutory controls, this results in a lower quality than would have been achieved on constrained data. Consequently, the analysis work underway will lead to an increase in data quality, first through the analysis of data where quality shortcomings are demonstrated, again on a proactive basis, and then through "friendly" alerts to entities for a short period, accompanied by proposals for corrections. Subsequently, binding controls will be implemented in the data acquisition phase, which, if not respected, will lead to the rejection of the data in the BDAP and to sanctions to entities not complying with the rules that have been defined by law and by administrative acts.An examination of the data showed that the lack of association of expenditure with COFOG differs according to the type of entity and also according to the items in the chart of accounts; the associations of COFOG with expenditure of a more typically financial nature are particularly lacking, such as the case of the repayment of loans, the closure of advances received from treasury banks and transactions on behalf of third parties.

7 Conclusions

The introduction of a reform of public accounting with the objective of controlling public finance extended to all entities or, at least, to its main groups, although of fundamental importance, will require a period of adaptation to the new rules that will be all the longer the more scattered the set of entities that make up the public administration sector is and the more a high level of professional training is required for the new system. Another fundamental issue to be considered concerning the control of public finance, both for expenditure and revenue, is based on the use of public budget data derived from the accounting management of the entities for the monitoring of programmatic objectives that are set on a more statistical representation. Consequently, on the one hand, it is necessary to train public accountancy professionals with a broader vision, including a statistical one, who can compare their own financial statements both in a time series and with those of similar bodies; on the other hand, it must be acknowledged, by the main users, especially those who operate from a more macroeconomic perspective and with a view to comparison with European objectives, that the accounting work of the bodies

needs time and respect so that the accounting reform fully meets its objectives through a process of study, presentation and sharing with the bodies of the analysis of their budgets and of the possibilities for qualitative improvement.

The purpose of a common representation that ensures the convergence of all functional classifications adopted by individual administrations, including the State, pays the tribute to the crucial role that the overall representation of public policies must play, both in the macroeconomic context of the whole public administration, as well as in a sectoral context that allows the analysis of expenditure and of the consequent results in terms of output and outcome related to the competences attributed to the different levels of government, in the context of public governance on the various levels of government, allowing the measurement of the activities for each of the actors (central or local) that contribute to the realisation of public policies.

References

1. https://www.rgs.mef.gov.it/_Documenti/VERSIONE-I/e-GOVERNME1/ARCONET/Glossari/Glossario-delle-missioni-e-dei-programmi-fino-al- 2024.pdf
2. Collesi, D.: “La classificazione della spesa per finalità economiche”, “Esame di alcuni casi pratici di rappresentazione della spesa per finalità”, “Enti e organismi stru- mentali: le missioni della spesa”, “Il glossario delle missioni e dei programmi per gli enti territoriali”, “Gli schemi di bilancio per Missioni e programmi degli enti territoriali”, publi- shed on Maggioli Review “La finanza locale”, several issues 2013–2014 (2013)

The Italian Statistics Olympiads

Patrizia Collesi(✉) and Antonio Pitrone

Italian National Institute of Statistics, 00184 Rome, Italy
{pcollesi,pitrone}@istat.it

Abstract. The paper presents the activities and results of the national competition for upper secondary students Italian Statistical Competition, started by the Italian Statistical Society in 2010 and jointly organized and managed with the Italian National Institute of Statistics – Istat since 2017. A historical overview of the activity is summarized analyzing the pages of the corporate websites (Sis and Istat) presenting it. Evolutions and changes are presented, including a fundamental addition to permit the participation to the European Statistics Competition which began in 2018. A second part is devoted to presenting participation data which shows the growth in participation, of both schools and students also at the territorial level, with a focus on students' evaluation.

Keywords: Italian statistics Olympics · Statistical education · Assessment

1 History and Evolution of the Italian Statistics Olympiads

1.1 The Project

The paper is the first attempt to present the history of the Italian Statistics Olympiads through the analysis of the documents available on the websites of the National Institute of Statistics (Istat) and the Italian Statistical Society (SIS) and is not exhaustive. Additional data could help to further analyze the competitions together with a more detailed assessment of participating students and their results. The Olympiads are a competition aimed at encouraging students to approach statistical reasoning, to arouse their interest in data analysis and probability, to enable them to better understand the meaning of the quantitative information they receive every day. They belong to the group of competitions for upper secondary schools, and are supported by the Italian Ministry of Education. They are part of the projects of Excellencies that permits winners to be inserted in the National Register of Excellencies.[1]

The competition is promoted and managed by the National Institute of Statistics (Istat) and the Italian Statistical Society (SIS).[2] Its main purpose is to disseminate statistical literacy with students, to foster the development of statistical thinking, the ability

[1] Patrizia Collesi authored paragraphs 1.2, 1.2; Antonio Pitrone authored paragraph 2.MIM, *Eccellenze-Normativa*, https://www.miur.gov.it/web/guest/tematiche-e-servizi/scuola/eccellenze/valorizzazione-delle-eccellenze/normativa.

[2] https://sis2024.sis-statistica.it/chi-siamo/

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 438–443, 2025.
https://doi.org/10.1007/978-3-031-64431-3_72

to interpret correctly quantitative synthesis, to arouse students' curiosity for data analysis and probability, and to encourage their interest towards the choice of statistics at university level.

1.2 The History and Evolution

The Italian Statistical Olympiads are the oldest uninterrupted initiative for school managed by Istat; it was actually began by the SIS in the 2010/2011 school year. Currently a history of the project does not exist, and the main sources to describe its evolution are SIS and Istat corporate websites.

In the first period the competition was intended for students attending the fourth year of upper secondary school, with two categories of participants. In the first group students attending specific courses of statistics were included, in the second one students studying statistics included in math courses.

In the 2011/2012 school year a first enlargement of possible users took place, consequently the Olympiads opened to the fourth and fifth year students. The purpose, as clearly stated in SIS website, was "to disseminate among the students approaching the choice of university choice a greater consciousness of the role of statistical sciences".

The third edition was held in March 2012, and the purposes and targets of the second edition were unchanged: i.e. to disseminate a wider awareness of the role and function of statistical sciences among the students of the last two years of the upper secondary school. Importance was given to the celebration of winners, that took place on the occasion of the Day of Statistics.

Also unchanged were the goals and target of the third edition (2014) and the fourth (2015).

Even unvaried, the 2016 edition became part of the "I Fuoriclasse della scuola" project, which rewards talented students with the financial contribution of trusts, associations, corporations in common activities. Students of the last three years of upper secondary schools with high scores in 13 individual competitions at the national level included in the Annual programme for the Excellencies are the recipients.[3]

The first relevant change was that of the 2017 edition (the seventh), when Istat entered as co-organizer in the Olympiads. The target were still students of the fourth and fifth years, with two tests: a qualifying test and a final one. Nevertheless a broader and more comprehensive goal emerged, not limited to foster the study of statistics, it was actually highlighted the purpose to "stimulate statistical literacy and the passion for the topics in an innovative, useful and engaging way".[4] In that same year the competition received the support of the National Project on Scientific Degrees-Statistics Sector, too. It was also broadened in scope, including students of the first and second year, that had a single test (without preliminary phase). Third and fourth years continued with both preliminary and final test, the only year that was left aside was the third.

For the first time the tests became available, with the publication on SIS website.[5] Starting from 2018 the solutions were available as well.

[3] SIS, *Progetto I fuoriclasse della scuola*, https://www.sis-statistica.it/ita/23/index.php?p=9905.

[4] Istat, *Al via le Olimpiadi di statistica del 2017*, 27 January 2017, https://www.istat.it/it/archivio/196010.

[5] SIS, *Prove anni passati*, https://www.sis-statistica.it/ita/23/index.php?p=6982.

Another relevant change was the addition of the team phase, to permit the participation of Italian teams to the European Statistics Competition (ESC), which had just started and which was one of the goals of the DIGICOM project of the European Commission for digital communication.

The eighth edition (2018) was the last including student of the fifth year, as well as the last including preliminary tests in the individual phase.

The ninth edition (2019) eliminated the preliminary tests and limited participation to fourth year student. Actually, it was experienced that fifth year students were strongly focused on the graduation exam and often registered for the competition but actually did not take part.

As from 2019 the situation is substantially stable as far as the organization is concerned. The competition has two phases: an on line individual test and a team phase. Students of the first four years of upper secondary school are admitted. They are grouped as follows: I-II years, III-IV years. Each group has the same test for the individual phase but the rankings are four (one per year). The test consists in 20 questions to be answered in 40 min on an online platform. The first three students per each year are inserted in the National Register of Excellencies. A second phase follows, that is the teams phase. Students participating in the individual phase can access if their school has obtained a given position in the school ranking. The team phase assignment consists in answering a research question on an official statistical topic with a presentation.

The following editions, from 2020 up to the current one (2024) have encountered no substantial revisions, with an overall remarkable steady growth of participants and schools. In 2024 Italian students confirms as the largest group of national students participating in the ESC ever, reaching nearly one third of participants (6639 out of 22479).

2 Participation Data Analysis

Over the years participation experienced a steady growing trend both for students and for popularity. It should be actually recalled that statistics is not an autonomous curricular topic in the last years of upper secondary school, and it has also to be added that the Statistics Olympiads should compare with other competitions dealing with curricular scientific topics such as mathemathics, physics, chemistry, which have a long-standing tradition, and for which the Italian phase is often a phase of the international competition.

As a way of example Italian students have taken part in Maths Olympiads since 1985, Physic Olympics began in 1987. The Chemistry Games became a national competition in 1987, and all of them are directly supported by the Ministry of Education.

Even not complete this overview shows how the participation of Italian students to the Statistics Olympiads is good. When comparing with the above-mentioned competitions it has to be added that the competition has only a European, not an international at large, phase. The European Statistics Competition actually started in 2018 and is not supported by any scientific society but by Eurostat only. Nevertheless, Italian students were the most numerous among all those from other EU countries over the years.

Figure 1 shows participation data starting from 2019, the first year for which comparable data are available concerning participating students: i.e. students from the first to the

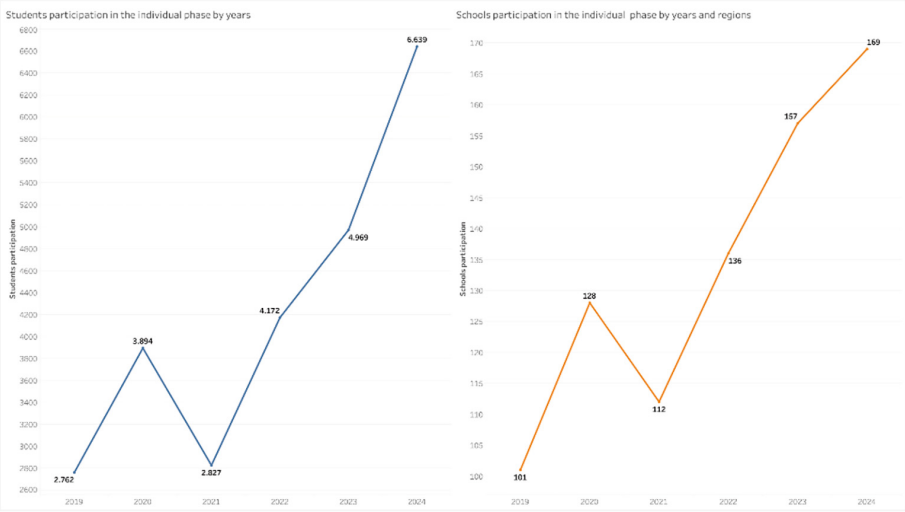


Fig. 1. Participation of students and schools by year (2019–2024)[6]

fourth year of upper secondary school. The participation rate grows almost ever evenly, with the exception of 2021 school year, a period of repeated openings and closures of schools due to the SARS-CoV-2 pandemic.

Analysing deeper the distribution of participation across the territory (Fig. 2), we can see that the highest number of participants, in terms of students and schools, were reached in Veneto and Emilia Romagna. Veneto passed from 560 students and 16 schools in 2019, to 1073 students and 20 schools in 2024; Emilia Romagna from 291 students and 13 schools in 2019, to 815 students and 18 schools in 2024.

The participation in Sicilia was quite constant through the six years analyzed, for both students' and schools' participation. Instead, Calabria had a constant growing trend along all the years despite the decrease of 2021–2022, due to the pandemic emergency. The trend is similar for Campania, although marked by a pause in 2023.

Moreover, the last two years show a good increase of Lombardia and Puglia. For the first region 112 students were attending, representing the 6 schools partecipating in 2019 while in 2024 the students were 784 from 18 schools. The second went from 136 students to 528 and from 5 to 15 schools.

Finally, the bottom maps in Fig. 2 presents the distribution of students' participation by year and province. It shows that the participation involves more and more territories throughout the country every year, becoming increasingly widespread.

Last but not least we focused our attention to the scores obtained by students by year, class and sex. Figure 3 shows the trend for four indicators concerning the evaluation in the individual phase of the competition: Maximum, Average, Minimum and Standard deviation.

[6] Every graph in this paper can be accessed at the link in references.

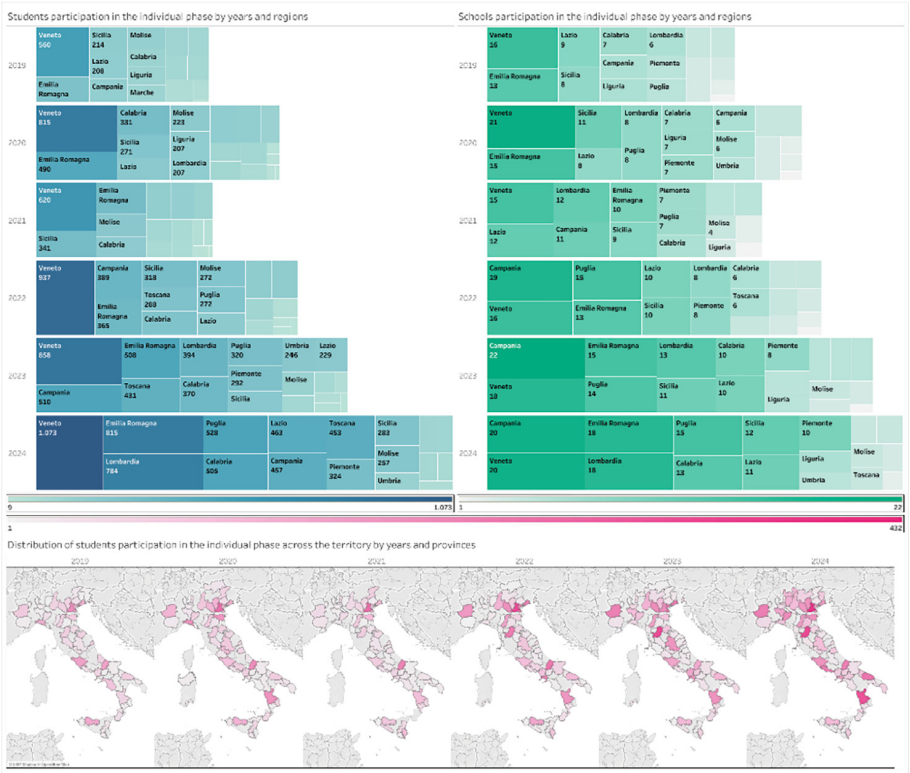


Fig. 2. Distribution of participation across the territory by year (2019–2024)

The high interval between minimum and maximum and the average value obtained by the students, together with the high standard deviation, show that much work needs to be done to disseminate statistical literacy.

Actually, the trend is encouraging. Net to the 2023 result, the average evaluation was always on the rise every year and for every class.

Class 1 and Class 2 students have always reached higher scores, this is probably due to the difference in educational programs and consequently in Statistical Olympiads test then.

It can also be observed that the gender difference is decreasing every year particularly in the first two classes. Probably also the reduction in the difference between gender evaluations in 2023 and 2024 for the last two classes, is to be attributed to the better job done with the younger female students in the previous years to bridge the gender gap.

This analysis is not exhaustive and can be further explored together by analyzing the quality of the questions in the test and the level of difficulty on the topics presented.

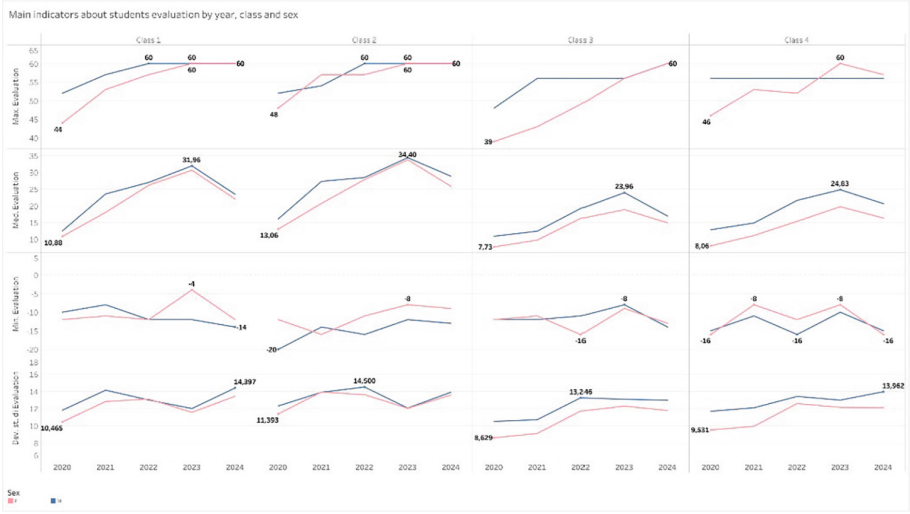


Fig. 3. Student evaluation by year, class and sex (2020–2024)

References

1. Barbieri, M.M., Pollice, A.: Le Olimpiadi Italiane di Statistica. Statistica e Società **2**, (2019)
2. Collesi, P., Mortati, F.M., Pitrone, A.: Il progetto digitale delle Olimpiadi Italiane di Statistica. Poster presented at XIV National Conference of Statistics. Rome (2021)
3. ESC: https://esc2022.eu/
4. ESC, Winning and Finalist Teams and Videos. https://www.esc2023.eu/
5. Italian National Institute of Statistics - Istat. "Le Olimpiadi italiane di statistica 2018". News, Istat, Rome, Italy (2018). https://www.istat.it/it/archivio/204605
6. Italian National Institute of Statistics - Istat. "Al via le Olimpiadi di statistica del 2017". News, Istat, Rome, Italy (2017). https://www.istat.it/it/archivio/196010
7. SIS, First edition Olimpiadi di Statistica. https://www.sis-statistica.it/ita/23/index.php?p=1663
8. SIS, First edition Olimpiadi di statistica. https://www.sis-statistica.it/ita/23/index.php?p=1668
9. SIS, Progetto I fuoriclasse della scuola. https://www.sis-statistica.it/ita/23/index.php?p=9905
10. SIS, Prove anni passati. https://www.sis-statistica.it/ita/23/index.php?p=6982
11. Collesi, P., Pitrone, A.: (2024) https://public.tableau.com/views/SIS2024-TheItalianStatisticsOlympiads/Participationofstudentsandschoolsbyyear2019-2024?:language=it-IT&publish=yes&:sid=&:display_count=n&:origin=viz_share_link



Learning Block Structures in Gaussian Graphical Models for Spectrometric Data Analysis

Alessandro Colombi[1(✉)], Lucia Paci[2], and Alessia Pini[2]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy
a.colombi10@campus.unimib.it

[2] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

Abstract. Probabilistic graphical modeling serves as a robust framework for capturing the conditional dependencies among variables that follow a Gaussian distribution. Within such models, each node represents a variable, and the absence of an edge between nodes indicates conditional independence given all other variables. Previous studies have applied this methodology to spectrometric data analysis, aiming at discovering the relationships among substances within a compound by analyzing their spectra. Such a goal has been achieved by coupling smoothing techniques for functional data analysis with a Bayesian Gaussian graphical model on basis expansion coefficients, hence simultaneously smoothing the data and providing an estimate of their conditional independence structure. Empirical evidence from real-world applications has shown that the adjacency matrix describing the underlying graph often presents a block structure. This implies a natural clustering of variables into disjoint groups. In this work, a new prior for Gaussian graphical models is introduced to learn the underlying clustering structure of the nodes. The method builds upon stochastic block models while accounting for the natural ordering of the nodes. The model is employed to analyze fruit purees and discover groups of portions of their spectra.

Keywords: Bayesian nonparametrics · Functional data analysis · Stochastic block models

1 Introduction

Infrared spectroscopy plays a crucial role in analyzing how infrared radiation interacts with matter through absorption, emission, or reflection. This technique is employed to study and identify chemical substances in solid, liquid,

L. Paci—Author 2 was partially supported by MUR-PRIN grant 2022 SMNNKY CUP J53D23003870008.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 444–449, 2025.
https://doi.org/10.1007/978-3-031-64431-3_73

or gaseous forms. Understanding the dependence structure among signals at different wavelength bands is essential for determining which bands are related to the different components. When two different bands of the spectrum are associated, it can be concluded that they pertain to either the same components or closely related ones. As infrared spectra are continuous functions of the wavelength, the analysis of spectrometric data can be interpreted as a smoothing problem of functional data, followed by inference on the smoothing coefficients. Using B-spline basis expansion to represent the functional data, each smoothing coefficient is associated with a specific basis function. Consequently, the vector of associated coefficients can depict the entire function. This approach enables the identification of hidden association between different bands of the spectrum by examining the relationships between the corresponding smoothing coefficients.

In order to model these dependence relations, [2] introduced Gaussian graphical models as a prior specification for the smoothing coefficients. These are probabilistic models that use undirected graphs to express the conditional dependence structure among variables through a joint Gaussian distribution. In this context, the nodes represent the smoothing coefficients, hence the graph interprets the dependence structure between the portions of the absorbance spectrum associated to such coefficients through the presence or absence of an edge between two nodes. The graph is unknown and must be learnt from the data. Following a Bayesian approach, the graph itself is considered a random variable, which can assume values in the space of all possible undirected graphs with p nodes.

In this work we focus on the analysis of a fruit purees dataset, already studied in [2,9]. Although using very different approaches, both works arrive at comparable graph structures characterized by significant associations between bands that are not closely located in the spectrum. In particular, the structure of dependencies is such that the adjacency matrix that describes the graph is block structured, i.e., connections happen mainly between groups of nodes.

The goal of this work is to make a step forward and simultaneously learn the graph and the clustering of the nodes that defines the block structure. Popular model-based clustering methods assume that objects to be clustered lack a specific ordering. However, in the case of spline coefficients, this assumption is not suitable since each node of the graph represents a band in the absorbance spectrum, and the bands are ordered according to the corresponding wavelengths.

To deliver model-based clustering, we rely on Stochastic Block Models (SBM) [4], a well-established class of models to infer the presence of communities in, usually, observed networks. In particular, we leverage recent works on SBM for latent networks [7,8] and exploit such modeling to introduce a prior for the graph that accounts for a random partition of the nodes while preserving their ordering constraints following [5]. We apply the method to discover clusters of spectrum bands of fruit purees, providing insights into potential relationships among substances within the compound.

2 Proposed Model

We consider an independent sample of n spectra, all evaluated at the same grid of r points, i.e., the wavelengths at which absorbance is measured, and model them using p basis functions. Let $y_i(s)$ be the absorbance spectrum of unit i at wavelength s. The model that we assume for the i-th spectrum is:

$$y_i(s) = \sum_{j=1}^{p} \beta_{ij}\varphi_j(s) + \varepsilon_i(s) \quad i = 1, \ldots, \text{ n}$$

where $\phi_j(s)$ is a cubic B-spline basis function, for $j = 1, \ldots, p$, $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^T$ is the spline coefficients' vector specific to i-th curve and $\varepsilon_i(s) \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau_\varepsilon^{-2}\right)$. We also introduce the cubic B-spline design matrix $\Phi \in \mathbb{R}^{r\times p}$ as $\Phi = [\varphi_j(s_k)]_{jk}$, where each element $\varphi_j(s_k)$ represents the j-th basis function evaluated at the k-th grid point s_k, for $j = 1, \ldots, p$ and $k = 1, \ldots, r$.

Let $\boldsymbol{Y}_i = (y_i(s_1), \ldots, y_i(s_r))^T$ be the absorbance spectrum at all observed wavelengths of curve i and assume

$$\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i, \tau_\varepsilon^2 \overset{\text{ind}}{\sim} \mathcal{N}_r\left(\boldsymbol{\Phi}\boldsymbol{\beta}_i, \mathbf{I}_r/\tau_\varepsilon^2\right) \quad i = 1, \ldots, n,$$

where τ_ε^2 is the error precision and $\mathbf{I}_r$ is the $r \times r$ identity matrix. We further assume $\tau_\varepsilon^2 \sim \text{gamma}(a_\varepsilon, b_\varepsilon)$. Following [2], we place a Gaussian graphical model as prior distribution for the basis expansion coefficients $\boldsymbol{\beta}_i$

$$\begin{aligned} \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n \mid \boldsymbol{\mu}, \boldsymbol{K} &\overset{\text{iid}}{\sim} \mathcal{N}_p\left(\boldsymbol{\mu}, \boldsymbol{K}^{-1}\right) \\ \boldsymbol{K} \mid G &\sim \text{G-Wishart}(b, D) \end{aligned}$$

where K is the precision matrix, $b > 2$ is the shape parameter and D is the $p \times p$ positive definite inverse scale matrix. Moreover, $\boldsymbol{\mu}$ is the prior mean vector, on which we place a Normal hyperprior, $\boldsymbol{\mu} \sim \mathcal{N}_p\left(0, \sigma_\mu^2 \mathbf{I}_p\right)$.

The graph G is random and belongs to the discrete space of all possible undirected graphs having p nodes. While most of the literature on graphical models places a Uniform or a Binomial distribution prior on G, we propose a new prior that accounts for a partition $\rho = (C_1, \ldots, C_M)$ of the p nodes in M mutually exclusive groups (or clusters), each with size $n_j = |C_j|$, for $j = 1, \ldots, M$. In general, the number of groups M is not known and must be learnt from the data. Moreover, given ρ, we also define $\boldsymbol{z} = (z_1, \ldots, z_p)$, a p-dimensional vector of cluster memberships defined through the following equivalence relationship: $i \in C_m \iff z_i = m$.

Due to ordering constraints, we follow [5] and consider only admissible partitions, i.e., partitions ρ such that, for each $i < j$, the label of each element of C_i is strictly smaller than the label of each element of C_j. As a prior for ρ, [5] presents a strategy to constraint an Exchangeable Partition Probability Function to the space of admissible partitions while preserving the induced prior on the number of clusters. In particular, we rely on a restriction of Ewens's sampling formula [3], with parameter θ.

Finally, the prior for the graph $G = (V, E)$ is defined through a SBM approach, i.e., the probability of having an edge between node i and node j only depends on their group membership z_i and z_j, respectively. Namely,

$$\mathbb{P}((i,j) \in E \mid \boldsymbol{z}, Q) \overset{\text{ind.}}{=} Q_{z_i,z_j}, \text{ with } i,j \in \{1,\dots,p\} \text{ and } i < j, \tag{1}$$

where Q is a $M \times M$ symmetric probability matrix such that $Q_{u,v}$ is the probability of having an edge between any node in cluster u and any other node in cluster v. Furthermore, we place i.i.d. Beta priors over the elements of $\boldsymbol{Q}$, $Q_{u,v} \mid M \overset{\text{iid}}{\sim} \text{Beta}(\alpha_1, \alpha_2)$, for $1 \leq u \leq v \leq M$. By exploiting the Beta-Bernoulli conjugacy, we integrate $\boldsymbol{Q}$ out of Equation (1). By doing so, the prior for the graph G, conditionally on the vector of group membership $\boldsymbol{z}$, is

$$\mathbb{P}(\boldsymbol{G} \mid \boldsymbol{z}) = \prod_{u=1}^{M} \prod_{v=u}^{M} \frac{\text{B}\left(\alpha_1 + S_{u,v}, \alpha_2 + S^{\star}_{u,v}\right)}{\text{B}(\alpha_1, \alpha_2)},$$

where $S_{u,v}$ is the number of existing edges between the nodes belonging to cluster u and the nodes in cluster v and $S^{\star}_{u,v}$ is the number of non-edges between the same two clusters.

Bayesian posterior inference is carried out through a block Gibbs sampling strategy. Firstly, we update the $\boldsymbol{\beta}$ coefficients, $\boldsymbol{\mu}$ and τ^2_{ε} from their full conditional distributions, which are available in closed form. Then, we perform a joint update of graph and precision matrix from their (joint) full conditional distribution. The latter is not available in closed analytical form and therefore we exploit the Birth-and-Death approach by [6]. Finally, the partition $\boldsymbol{z}$ is updated using an adaptive Split-and-Merge algorithm, as in [1]. Moreover, to improve the mixing of the chain a shuffle move is performed.

3 Analysis of Fruit Purees Dataset

The motivating problem is the analysis of spectrometric data of fruit purees. The analyzed dataset consists on a sample of $n = 351$ strawberry purees measured on an equally spaced grid of 235 different wavelengths on the interval $I = [899.327\,\text{nm}, 1802.564\,\text{nm}]$. The problem of interest is in determining the chemical composition of the purees samples. In particular, if different regions of the spectrum are related, then they might correspond to the same chemical component. Thanks to the compact supports of the B-spline basis, which identify well-defined portions of the domain, the problem of studying interactions between different substances simply translates into studying the dependencies between the basis expansion coefficients. Using the Gaussian graphical model presented in Sect. 2, the latter can be read off from the graph G. We fit the model with $p = 20$ basis functions. We decided to use a limited number of bases to not overfit the data. Indeed, the shape of the spectra is regular and similar for all curves: they all exhibit a large initial peak, a flat portion in the middle and two secondary peaks in the tail of the spectrum. As in [2], we set $b = 3$,

$D = \mathbf{I}_r$, $a_\varepsilon = 10$, $b_\varepsilon = 0.001$, $\sigma_\mu^2 = 100$. Then, the SBM prior is completed by setting $\alpha_1 = 12$ and $\alpha_2 = 3$, so that the expected density of each block of the adjacency matrix is 0.8. Finally, the concentration parameter θ is random with hyperprior $\theta \sim \text{gamma}(1, 1)$. The total number of Gibbs sampler iterations is $100,000$, of which $10,000$ iterations were discarded as a burn-in.

The smoothed curves, as well as the estimated partition of the spectrum, are depicted in Fig. 1, while the corresponding estimated graph is reported in the left panel of Fig. 2. The right panel of Fig. 2 depicts the posterior distribution of the number of groups, where 4 is the posterior mode but also 5 is strongly supported by the data. Note that almost all the posterior mass is far from 1, that is when all nodes belong to the same group and therefore the model collapses to the one presented in [2]. The estimated graph shows a comparable graph structure both with the findings of [2,9]. The first group represents the part of the spectrum before the main peak. This is not dense in terms of connections within the group but it is instead connected both with nodes in the second and last group. The second group is the largest, as it includes the main peak. This is the most informative part of the signal, and - from previous analyses - we know it has many short-term and long-term dependencies. As expected, this group has many connections within the group, and is also connected with all other groups. The third one is the flattest portion of the spectrum which is the only block of the graph with no intra-group connections. Finally, the final group include the two final peaks in the tail of the spectrum, each of which is connected with the first two groups.

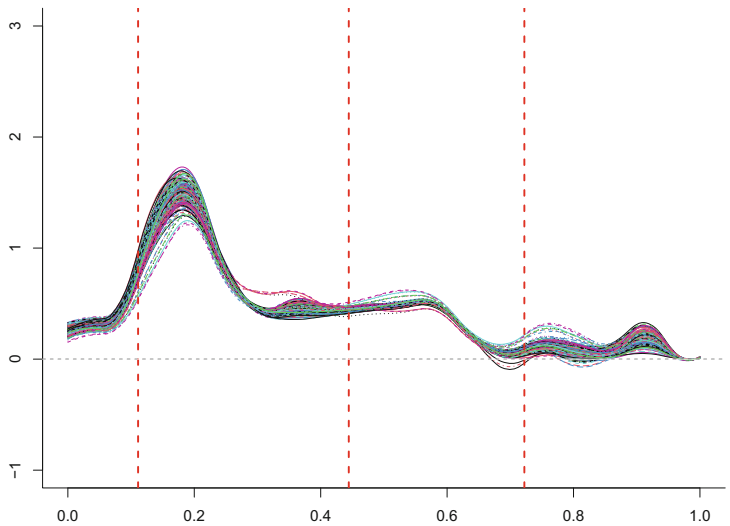


Fig. 1. Plot of the 351 smoothed curves along with the estimated partition of the spectra. Groups are delimited using vertical red dotted lines.

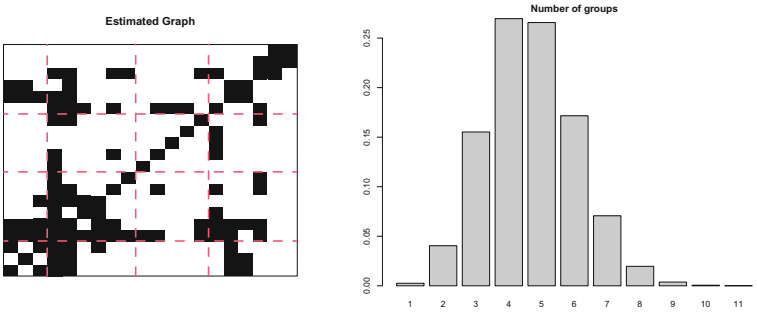


Fig. 2. Left: posterior graph estimate; Right: posterior distribution of the number of groups.

References

1. Benson, A., Friel, N.: Adaptive MCMC for multiple changepoint analysis with applications to large datasets. Electron. J. Stat. **12**(2), 3365–3396 (2018)
2. Codazzi, L., Colombi, A., Gianella, M., Argiento, R., Paci, L., Pini, A.: Gaussian graphical modeling for spectrometric data analysis. Comput. Stat. Data Anal. **174**, 107416 (2022)
3. Ewens, W.J.: The sampling theory of selectively natural allels. Theor. Popul. Biol. **3**, 87–112 (1972)
4. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)
5. Martínez, A.F., Mena, R.H.: On a nonparametric change point detection model in Markovian regimes. Bayesian Anal. **9**(4), 823–857 (2014)
6. Mohammadi, A., Wit, E.C.: Bayesian structure learning in sparse Gaussian graphical models. Bayesian Anal. **10**(1), 109–138 (2015)
7. Tan, L.S.L., Jasra, A., De Iorio, M., Ebbels, T.M.D: Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. Ann. Appl. Stat. **11**(4), 2222–2251 (2017)
8. van den Boom, W., De Iorio, M., Beskos, A.: Bayesian learning of graph substructures. Bayesian Anal. **18**(4), 1311–1339 (2023)
9. Waghmare, K.G., Panaretos, V.M.: Continuously indexed graphical models. arXiv:2302.02482 (2023)



On the Evidence of a New Bayesian Discrepancy Measure for Discrete Scalar Parameters

Silvia Columbu(✉), Francesco Bertolino, and Monica Musio

University of Cagliari, Cagliari 09120, Italy
silvia.columbu@unica.it

Abstract. We explore the applicability of the Bayesian Discrepancy Measure, a measure of evidence recently introduced in the literature, in the case of models involving hypotheses on discrete parameters. The approach is flexible, and can be adapted to take into account different distributions also when such non-regularity conditions are met. We investigate more in details the case of hypotheses on the population size parameter for capture-recapture models. We present simulation studies and a data application.

Keywords: BDM · Discrete parameters · capture-recapture

1 Introduction

The Bayesian Discrepancy Measure (BDM) is a Bayesian evidence measure recently introduced in the literature ([4]) that serves as a tool for evaluating precise hypotheses. It has been demonstrated that this measure, given certain regularity conditions, remains invariant under reparametrizations and exhibits asymptotic consistency. Initially defined for univariate models, the BDM has also been extended to compare parameters across two independent populations (see [6]). A recent study [5] delved into comparing coefficients of variation under various model assumptions. Notably, one of the principal advantages of the BDM lies in its applicability, without the imposition of stringent conditions on involved models and parameters, once identifiability is met. In this contribution we want to highlight the versatility and effectiveness of the BDM when applied to non-regular models, to tackle issues and scenarios that have received insufficient attention in the literature. Our focus centers on discrete parameters, with a specific examination of the capture-recapture model for closed populations, as a case study. We provide evidence, in case of scalar parameters, of the validity of the method for capture-recapture model, presenting simulation studies and an application to real data.

2 The BDM for Regular Models

Let $X \sim f(x|\theta)$ be a parametric model indexed by a scalar parameter $\theta \in \Theta \subseteq \mathbb{R}$, $g_0(\theta)$ a prior density, $\mathbf{x}$ a sample of iid observations and $g_1(\theta) \propto g_0(\theta)\ L(\theta|\mathbf{x})$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 450–456, 2025.
https://doi.org/10.1007/978-3-031-64431-3_74

a posterior density, where $L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ is the likelihood function. Let also $G_1(\theta|\mathbf{x})$ be the correspondent distribution function and $m_1 = G_1^{-1}\left(\frac{1}{2}|\mathbf{x}\right)$ the posterior median, supposed to be unique. We are interested in evaluating sharp hypotheses such as $H : \theta = \theta_H$. The definition of BDM (δ_H) is based on the discrepancy of the hypothesis from the posterior median, in particular

$$\delta_H = 1 - 2 \cdot \int_{I_E} dG_1(\theta|\mathbf{x}) = 1 - 2 \cdot \min\left\{P(\theta > \theta_H|\mathbf{x}),\, P(\theta < \theta_H|\mathbf{x})\right\}, \quad (1)$$

where I_E is the external interval expressed as

$$I_E = \begin{cases} (\theta_H, +\infty) \text{ if } & \theta_H > m_1 \\ (-\infty, \theta_H) \text{ if } & \theta_H < m_1. \end{cases}$$

The value of δ_H is comprised between 0 and 1, approaching the upper bound whenever θ_H is farther from the median of the posterior distribution $G_1(\theta|\mathbf{x})$, in this case it can be said that H does not conform to $G_1(\theta|\mathbf{x})$. Values closer to 0 support the evidence in favor of H.

This definition has been given under standard regularity conditions. In [4] it has been proved that, in case of a scalar parameter, the following properties hold:

(i) δ_H always exists and, by construction, $\delta_H \in [0, 1)$;
(ii) δ_H is invariant under invertible monotonic transformations of the parameter;
(iii) if θ is an a.c. random variable, θ^* is the true value of the parameter and $\theta^* = \theta_H$, then δ_H converges asymptotically to a $Unif(\cdot|0; 1)$. Otherwise, if $\theta^* \neq \theta_H$, then $\delta_H \xrightarrow{p} 1$ (consistency property).

3 The BDM for Discrete Parameters

When dealing with non-regular models, in particular models that rely on non continuous parameters or in which the support depends on the parameter of interest, it is still possible to apply the BDM, albeit with caution. The approach must be tailored to accommodate the specific characteristics of the problem arising from the irregularity of the model. For instance, if one intends to use non-informative prior laws, it's advisable to steer clear of Jeffreys' priors, which rely on expected Fisher information. Instead, one might consider employing reference priors, as extensively discussed by Bernardo and Smith [3].

When the parameters of interest are discrete, with $\theta \in \mathbb{Z}$, it is possible to determine the exact probability of the hypothesis, $\mathbb{P}\big(\theta = \theta_H|\boldsymbol{x}\big) > 0$. Notably, the formulation in equation (1) necessitates adaptation as follows:

$$\delta_H = 1 - 2 \cdot \min\left\{P(\theta > \theta_H|\boldsymbol{x}),\, P(\theta < \theta_H|\boldsymbol{x})\right\} - P(\theta = \theta_H|\boldsymbol{x}). \quad (2)$$

3.1 BDM for Capture-Recapture Models

Capture-recapture methods originally arised in the ecology field for the estimation of wildlife population. Over time, their applicability has expanded to various fields requiring abundance estimation in closed populations, including epidemiology, demography, and official statistics, among others. The classical two-sample capture-recapture model, introduced by Petersen in 1896 ([7]), relies on the hypergeometric distribution. The estimation of the BDM presented in this work involves revisiting this classical model.

Consider a population of size N and two sampling occasions (captures) of fixed sizes n_1 and n_2, where s denotes the number of units (recaptures) common to both samples. The probability of the number of recaptures follows the hypergeometric distribution, and can be used to infer the value of $N \in \mathcal{N} = r, r+1, \ldots$, where $r = n_1 + n_2 - s$ represents the total number of units observed in the two samplings. We aim to apply the discrete version of δ_H for testing the hypothesis.

$$H : N = N_H$$

with N_H a specific value in $\mathcal{N}$. Under the hypergeometric model the likelihood for the parameter N is

$$L(N|n_1, n_2, s) \propto \ Hyp(s|n_2, N - n_2, n_1) = \frac{\binom{n_2}{s}\binom{N-n_2}{n_1-s}}{\binom{N}{n_1}} \cdot \mathbf{1}_{\mathcal{N}}(N).$$

Given the discreteness of the parameter space, the choice of an objective prior distribution is a critical point. It is in fact known that the usual choice of objective prior for a discrete parameter is the constant prior (for a discussion on objective priors for discrete parameters see [2]). Other possible choices concern the application of reference priors, in particular we propose to assume as prior distribution $g_0(N) \propto \ N^{-\frac{3}{2}}$ which ensures a correct decay of the associated posterior

$$g_1(N|n_1, n_2, s) \propto \frac{1}{N^{\frac{3}{2}}} \cdot \frac{\binom{n_2}{s}\binom{N-n_2}{n_1-s}}{\binom{N}{n_1}} \cdot \mathbf{1}_{\mathcal{N}}(N) \ .$$

Given $g_1(N|n_1, n_2, s)$, the discrepancy δ_H can be computed as in (2) with

$$\mathbb{P}\big(N < N_H\big) = \sum_{i=r}^{N_H-1} g_1(i); \ \mathbb{P}\big(N = N_H\big) = g_1(N_H); \ \mathbb{P}\big(N > N_H\big) = \sum_{i=N_H+1}^{+\infty} g_1(i).$$

4 Empirical Analysis

4.1 Simulations

In the case of capture-recapture models it is not clear if property (iii) of Sect. 2 still holds, since the proposed proof hinges on the absolute continuity of the

parameter under study (see [4]). We perform simulation studies to show that, with increasing sample and populations sizes, the property still persists. This could be not surprising as it is well known that, for large N, the hypergeometric distribution converges to the binomial one with parameter $p = s/N$. Therefore, in order to verify the asymptotic consistency of δ_H and investigate its distribution, we generated recaptured observations from the hypergeometric distribution letting vary the sample sizes of the two drawnings as well as the total population size. We observe that when data is generated under the hypothesis H, that is $N = N_H$, then, for large N, H is likely to fall almost uniformly anywhere between 0 and 1. However, with small samples and small populations, we observe a discretness of δ_H distribution. This result is in agreement with the nature of the parameter distribution (see Fig. 1). Simulations suggest also that the asymptotic consistency still remains: the proportion of δ_H values close to 1 is higher whenever the hypothesis N_H does not hold, and in particular is almost 1 if the distance from N increases (see Table 1).

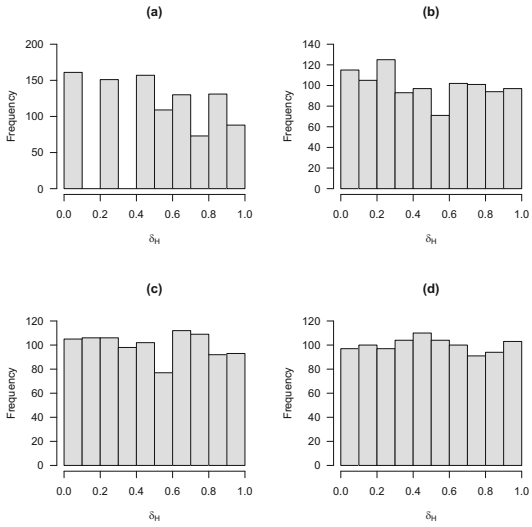


Fig. 1. Distribution of δ_H values for data generated under the hypothesis $N = N_H$ for increasing sizes. The following model parameters were set: (a) $n_1 = 30$, $n_2 = 50$, $N = N_H = 100$; (b) $n_1 = 450, n_2 = 300, N = N_H = 1000$; (c) $n_1 = 5000, n_2 = 4000, N = N_H = 10000$; (d) $n_1 = 40000, n_2 = 50000, N = N_H = 100000$. A total of 1000 replicates were considered in all settings.

Table 1. Proportion of δ_H values under two different settings of hypergeometric model, with increasing sizes, over 1000 replications.

	$N_H = 70$	$N_H = 100$	$N_H = 130$	$N_H = 160$
$n_1 = 30, n_2 = 50, N = 100$	0.97893	0.50167	0.71261	0.90620
	$N_H = 700$	$N_H = 1000$	$N_H = 1300$	$N_H = 1600$
$n_1 = 450, n_2 = 300, N = 1000$	1	0.50285	0.99796	0.99999

4.2 Data Application

In Western countries cancer is the primary cause of death, therefore the implementation of epidemiological programs for prevention and control becomes of crucial importance. The success of such programs is subject to the availability of reliable and complete data, such as cancer registers and screening programs. Indeed, the application of capture-recapture models can help in the evaluation of the representativeness of data collection systems with respect to the real population. In this direction, Bailly et al. [1] proposed a bayesian capture-recapture model whose parametrization takes into account the individual capture heterogeneity. They tested their method in the estimation of colorectal cancer populations in Southeastern France, taking as capture-recapture episodes data collected from the Histopathological Registry and the hospital Multidisciplinary Team Meetings in Alpes Maritimes in 2008. The total number of cases observed is $N = 509$, where $n_1 = 481$ come from the registry, $n_2 = 328$ from the hospital team and $s = 300$ are the recaptured units common to both sources. Their method provided an estimation of 526.5 cases, with associated an HPD Interval of $(519 - 535)$. We apply the BDM to evaluate the conformity of this result to the data observed. The hypothesis to test is $N = N_H = 526$ and the associated BDM is $\delta_H = 0.0268$. This small value suggests that there is not enough evidence against the hypothesis made, in fact N_H is almost coincident with the median of the posterior density. The result is graphically summarized in Fig. 2 where it is reported the posterior density g_1 of the population size parameter, together with the value N_H under investigation. In some sense, in this application our method can be taken as a tool for the validation of the goodness of the estimation model in [1].

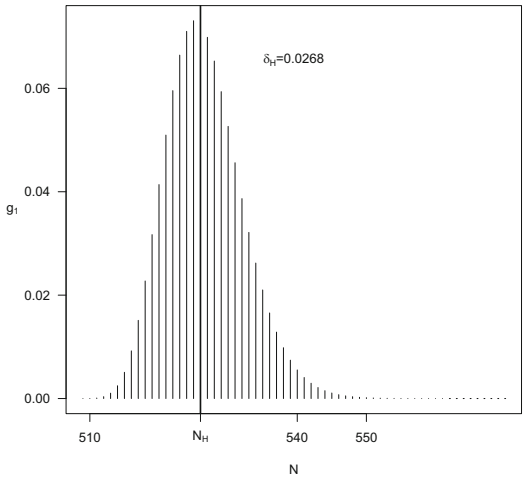


Fig. 2. Posterior density in the colorectal cancer application.

5 Conclusions

In this contribution we have proposed to extend the BDM to evaluate hypotheses concerning discrete scalar parameters in capture-recapture models. The method presented benefits of a simple and clear definition and allows to take into consideration statistical frameworks which are of interest in many applications. This is the first contribution in which the BDM is proposed to test hypotheses in non-regular models, the extension to other cases is under investigation. Specifically, we are exploring cases where the hypothesis H entails parameters that appear in the support of the model.

References

1. Bailly, L., Daurès, J.P., Dunais, B., et al.: Bayesian estimation of a cancer population by capture-recapture with individual capture heterogeneity and small sample. BMC Med. Res. Methodol. **15**, 39 (2015)
2. Berger, J.O., Bernardo, J.M., Sun, D.: Objective priors for discrete parameter spaces. J. Am. Stat. Assoc. **107**(498), 636–648 (2012)
3. Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, NY (1994)
4. Bertolino, F., Manca, M., Musio, M., Racugno, W., Ventura, L.: A new Bayesian discrepancy measure. Stat. Methods Appl. **33**, 381–405 (2024). https://doi.org/10.1007/s10260-024-00745-1
5. Bertolino, F., Columbu, S., Manca, M., Musio, M.: Comparison of two coefficients of variation: a new Bayesian approach. Commun. Stat. Simul. Comput. (2023). https://doi.org/10.1080/03610918.2023.2231179

6. Manca M., Columbu S., Musio M.: A Bayesian test for the comparison of two independent populations. In: Proceedings of the 51th Italian Statistical Society Scientific Meeting (2022)
7. Petersen, C.G.J.: The yearly immigration of young plaice into the Limfjord from the German sea. Rep. Dan. Biol. Stat. **6**, 5–84 (1896)

A Direct Estimate of Lorenz Curve and Inequality Measures in European Countries

Francesca Condino(✉)

Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Rende, CS, Italy
francesca.condino@unical.it

Abstract. In this work, the attention is focused on the possibility to directly estimate Lorenz curve and income inequality characteristics by using a new approach, involving Lorenz curve derivative, the so-called share density. In a parametric perspective, the maximum likelihood method is used to obtain parameter estimates characterizing the share density, by considering in a proper manner the observed data. Preliminary results regarding 27 different European countries are obtained, starting from European Union Statistics on Income and Living Conditions (EU-SILC) cross-sectional data.

Keywords: Gini coefficient · maximum likelihood estimation · tail inequality

1 Introduction

The Lorenz curve is a well-known and widely used tool for analysing income inequality. From its proposal in 1905 [7], its study has generated a lot of investigation and a fertile field of research for economists and statisticians, to the present day. In dynamic clustering context, a recent paper [1], put the attention on Lorenz density, a tool much less explicitly mentioned than Lorenz curve, but equally informative. Actually, the Lorenz density, also called share density, furnishes different information regarding income inequality and associated measures [9]. In this work, starting from this particular function and its interpretation, the possibility to describe income inequality in different European countries through a direct estimate of the Lorenz curve is explored. To this end, EU-SILC data are considered.

2 The Parametric Approach

Here, a parametric approach to describe income distribution and its characteristics is considered, although many different methods are available. Therefore, the choice of an a priori functional form for income data description is required. We assume that Y is a non-negative continuous random variable (rv), describing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 457–462, 2025.
https://doi.org/10.1007/978-3-031-64431-3_75

income, with distribution function (df) $F(y;\boldsymbol{\theta})$ and probability density function (pdf) $f(y;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. Given a such parametric model, it is possible to obtain the expression of the associated parametric Lorenz curve and its derivative, as follows:

$$\eta(u;\boldsymbol{\theta}) = \frac{\int_0^u F^{-1}(t;\boldsymbol{\theta})dt}{E(Y;\boldsymbol{\theta})} \tag{1}$$

and

$$\eta'(u;\boldsymbol{\theta}) = \frac{\partial \eta(u;\boldsymbol{\theta})}{\partial u} = \frac{F^{-1}(u;\boldsymbol{\theta})}{E(Y;\boldsymbol{\theta})}. \tag{2}$$

where $F^{-1}(t;\boldsymbol{\theta}) = \sup\{y|F(y;\boldsymbol{\theta}) \leq t\}$ is the quantile function and $E(Y;\boldsymbol{\theta}) = \int_0^1 F^{-1}(t;\boldsymbol{\theta})dt$ is the expectation of Y. It is well known (see, e.g. [6]) that the Lorenz curve $\eta(u;\boldsymbol{\theta})$ can be viewed as a distribution function of a U rv, having support in the unit interval and, as a consequence, the corresponding derivative with respect to u, reported in (2), can be considered a density function. Known the distribution function of Y, it is often possible to identify the distribution function of U and its properties. Different examples are reported in [1,8], where the expression of Lorenz curves associated with some classical income distributions are furnished. Moreover, as highlighted by [4], U is a rv related to a particular experiment in which, for each unit of amount, chosen at random, the specific percentile range of the person who earned that unit is recorded. Therefore, given the observed data, a direct estimate of the parameters characterizing the density in (2) can be obtained by applying the usual methods, such as the maximum likelihood method.

3 Some Results for the European Countries

To illustrate this proposal, cross-sectional data from European Union Statistics on Income and Living Condition (EU-SILC) dataset are used. The survey aims to collect information on income, poverty, social exclusion, and living conditions. In particular, the disposable equivalent household income is considered, to take into account the real composition of the households. We assume that the income follows a Dagum distribution [2,5], since this model generally provides excellent fit to these kind of data. Therefore, the df is $F_{Da}(y;\beta,\lambda,\delta) = (1+\lambda y^{-\delta})^{-\beta}$ where $\lambda > 0$ is a scale parameter and $\beta, \delta > 0$ are shape parameters and inverse indicators of inequality [3]. We remember that, for this model, the $r-th$ moments exists whenever $\delta > r$. Under these hypotheses, U follows a generalized beta of the first type (GB1) [1], i.e. $U \sim GB1(\frac{1}{\beta}, 1, \beta + \frac{1}{\delta}, 1 - \frac{1}{\delta})$, with:

$$\eta(u;\boldsymbol{\theta}) = \frac{B(u^{1/\beta};\beta + \frac{1}{\delta}, 1 - \frac{1}{\delta})}{B(\beta + \frac{1}{\delta}, 1 - \frac{1}{\delta})} \tag{3}$$

where $B(x;.,.)$, with $x \in (0,1)$ is the incomplete Beta function. Let us consider the available observations on income and arrange them in increasing order, $y_{1:n}$,

$y_{2:n},...,y_{n:n}$, for $i = 1,..,n$. If we focus on the probability that a single unit of amount, chosen at random, is earned by a specific range of population, under the hypothesis of independence between the units of amount and by using the expression (3), after some passages, it is possible to obtain the expression of the likelihood function. This turn out to be a weighted sum, where each weight is defined by the amount $y_{i:n}$ earned by the corresponding range of population $(u_{i-1}, u_i] = \left(\frac{i-1}{n}, \frac{i}{n}\right]$, i.e.:

Table 1. Fitted parameters, corresponding standard errors (in brackets) and inequality indices values for the European countries.

	$\hat{\beta}$ (St. dev.)	$\hat{\delta}$ (St. dev.)	G_{obs}	$G(\hat{\delta}, \hat{\beta})$	T_{obs}	$T(\hat{\delta}, \hat{\beta})$
Austria	0.835 (0.0006)	3.544 (0.0007)	0.2845	0.2954	0.1576	0.1546
Belgium	1.333 (0.0012)	3.502 (0.0007)	0.2629	0.2686	0.1323	0.1308
Bulgaria	1.057 (0.0015)	2.631 (0.0007)	0.3804	0.3760	0.2732	0.2723
Cyprus	1.600 (0.0039)	2.967 (0.0009)	0.3104	0.3092	0.1821	0.1814
Czechia	1.335 (0.0019)	3.765 (0.0011)	0.2528	0.2492	0.1117	0.1113
Germany	0.907 (0.0003)	3.475 (0.0003)	0.2925	0.2948	0.1557	0.1550
Denmark	1.665 (0.0025)	3.381 (0.0007)	0.2612	0.2680	0.1340	0.1325
Estonia	0.504 (0.0004)	3.922 (0.0014)	0.3330	0.3156	0.1784	0.1711
Greece	0.822 (0.0007)	3.419 (0.0009)	0.3034	0.3072	0.1685	0.1682
Spain	0.821 (0.0007)	3.421 (0.0009)	0.3160	0.3071	0.1685	0.1680
Finland	11.051 (0.0015)	2.594 (0.0002)	0.2922	0.3134	0.2236	0.2046
France	1.081 (0.0006)	3.509 (0.0004)	0.2737	0.2799	0.1413	0.1407
Croatia	0.537 (0.0006)	3.878 (0.0018)	0.3160	0.3116	0.1678	0.1673
Hungary	1.344 (0.0026)	3.348 (0.0012)	0.2685	0.2809	0.1472	0.1444
Ireland	2.335 (0.0015)	2.481 (0.0002)	0.3472	0.3571	0.2648	0.2606
Italy	0.818 (0.0003)	3.050 (0.0003)	0.3358	0.3438	0.2176	0.2156
Lithuania	0.914 (0.0009)	2.842 (0.0007)	0.3637	0.3588	0.2411	0.2404
Luxembourg	0.534 (0.0003)	4.181 (0.0010)	0.2917	0.2914	0.1456	0.1451
Latvia	0.642 (0.0006)	3.150 (0.0009)	0.3605	0.3563	0.2279	0.2275
Malta	1.154 (0.0015)	3.120 (0.0009)	0.3113	0.3109	0.1790	0.1783
Netherlands	1.277 (0.0010)	3.552 (0.0006)	0.2605	0.2669	0.1300	0.1287
Poland	0.798 (0.0006)	3.871 (0.0009)	0.2754	0.2745	0.1317	0.1315
Portugal	0.827 (0.0005)	3.088 (0.0005)	0.3432	0.3388	0.2099	0.2089
Romania	0.449 (0.0007)	4.333 (0.0035)	0.3007	0.3020	0.1554	0.1548
Sweden	0.943 (0.0006)	3.660 (0.0007)	0.2665	0.2773	0.1422	0.1362
Slovenia	0.959 (0.0009)	3.969 (0.0011)	0.2540	0.2546	0.1138	0.1136
Slovakia	0.552 (0.0008)	5.646 (0.0040)	0.2180	0.2176	0.0797	0.0794

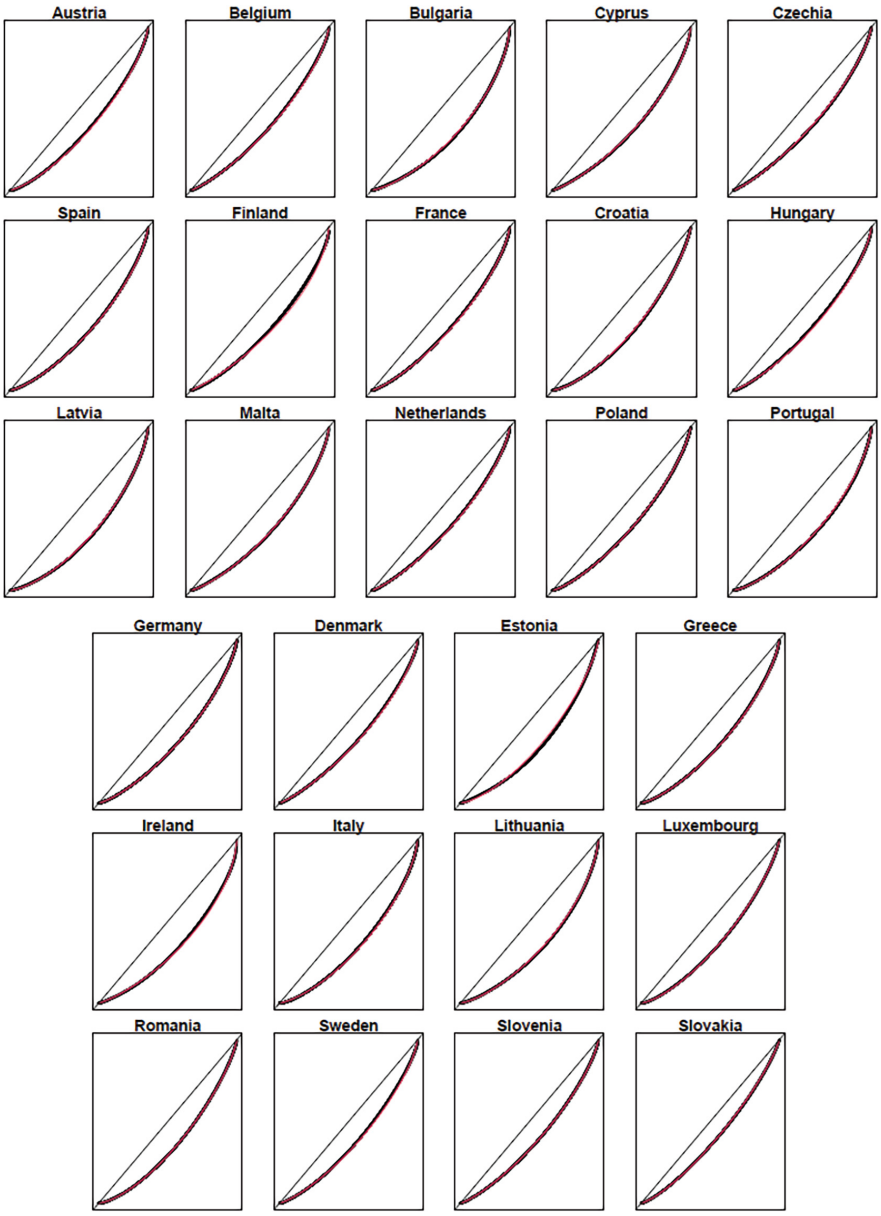


Fig. 1. Empirical (in black) and fitted (in red) Lorenz curve for 27 European countries. The 45-degree diagonal line is representative of perfect equality.

$$l(\boldsymbol{\theta};\boldsymbol{u}) \simeq \sum_{i=1}^{n} y_{i:n} \cdot \log \left[\frac{B\left(u_i^{1/\beta};\beta+\frac{1}{\delta},1-\frac{1}{\delta}\right) - B([u_{i-1}^{1/\beta};\beta+\frac{1}{\delta},1-\frac{1}{\delta})}{B(\beta+\frac{1}{\delta},1-\frac{1}{\delta})} \right]. \quad (4)$$

By optimizing the previous function, we obtain the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\delta})$ of $\boldsymbol{\theta}$. Here, we consider the ordered observed income for 27 different European countries $y^{(j)}_{1:n_j}, y^{(j)}_{2:n_j}, ..., y^{(j)}_{n_j:n_j}$ $(j = 1, ..., 27)$, and, contextually, the observed cumulative proportions of population $\frac{1}{n_j}, \frac{2}{n_j}, ..., 1$. Using the expression (4), the MLEs are obtained for each country. Results are reported in Table 1. From the comparison between empirical and fitted Lorenz curves shown in Fig. 1, it is evident the ability of the proposed method to describe the observed income concentration. Moreover, to further evaluate the accuracy of this approach in properly describe particular aspects of inequality, a comparison between observed and fitted values for the two widely used inequality indices, Gini (G) and Theil's (T), is shown. The findings, reported in Table 1, suggest very close fits almost everywhere and a good performance of this approach in describing the specific inequality aspects for each country.

4 Conclusions

Among all the possible ways for drawing knowledge from the Lorenz curve, in this paper we propose to focus on its derivative, the so-called share density, a particular function that provides a good deal of information regarding inequality. Considering in a proper manner the observed data regarding the amount earned by the different households, we propose to estimate the parameters characterizing the Lorenz curve, using a weighted maximum likelihood function. To this end, we start from the hypothesis that income follows a Dagum distribution in order to obtain the expression of the associated Lorenz curve and, consequently, the likelihood's expression. The obtained results for 27 different European countries clearly show the ability of this method to properly fit the observed curves in very different contexts, such as those represented by the various countries. This evidence is also confirmed by the comparisons between fitted and observed Gini and Theil's inequality indices that appear very close, both in condition of low and high inequality.

Funding Information. The paper's empirical part is based on data from Eurostat, EU Statistics on Income and Living Conditions survey (2023). The responsibility for all conclusions drawn from the data lies entirely with the author.

Data Availability Statement. The author gratefully acknowledges research grant 'Fondo sostegno aree socio-umanistiche-Cda del 26.03.2021-quota DESF from the University of Calabria.

References

1. Condino, F.: Share density-based clustering of income data. Stat. Anal. Data Min. ASA Data Sci. J. **16**(4), 336–347 (2023). https://doi.org/10.1002/sam.11619
2. Dagum, C.: A new model of personal distribution: specification and estimation. Econ. Appl. **30**, 413–437 (1977)

3. Dancelli, L.: Confronti fra le curve di concentrazione Z(p) e L(p) nel modello di Dagum. Stat. Appl. **1**, 399–415 (1989)
4. Farris, F.A.: The Gini index and measures of inequality. Amer. Math. Mon. **117**(10), 851–864 (2010)
5. Kleiber, C.: A guide to the Dagum distribution. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves Series: Economic Studies in Inequality, Social Exclusion and Well-Being (2008)
6. Kleiber, C.: The Lorenz curve in economics and econometrics. In: Betti, G., Lemmi, A. (eds.) Advances on Income Inequality and Concentration Measures: Collected Papers in Memory of Corrado Gini and Max O. Lorenz. Routledge, London (2008)
7. Lorenz, M.O.: Methods of measuring the concentration of wealth. Publ. Am. Stat. Assoc. **9**(70), 209–219 (1905)
8. Sarabia, J.M.: Parametric Lorenz curves: models and applications. In Chotikapanich D. (eds.) Modeling Income Distributions and Lorenz Curve: Economic Studies in Equality, Social Exclusion and Well-Being, vol. 5, Springer, New York (2008)
9. Rohde, N.: Lorenz curves and generalised entropy inequality measures. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curve: Economic Studies in Equality, Social Exclusion and Well-Being, vol. 5, Springer, New York (2008)

A Kruskal-Wallis Type Test for Multi-treatment Effect

Pier Luigi Conti[1] and Livia De Giovanni[2(✉)]

[1] Sapienza Università di Roma, Roma, Italy
pierluigi.conti@uniroma1.it
[2] LUISS, Roma, Italy
ldegiovanni@luiss.it

Abstract. The problem of testing for the presence/absence of a treatment effect in case of multi-treatment is attacked here through a new test-statistic based on the same ideas as the classical Kruskal-Wallis test. Its theoretical properties are studied, under both the null and the alternative hypothesis, and an approximation based on sub-sampling is proposed. Then, the problem of testing for stochastic dominance is also dealt with.

Keywords: Multi-treatment · Nonparametrics · Kruskal-Wallis · Stochastic dominance

1 Introductory Aspects

Assessing whether a treatment has effect on an outcome is of primary interest in Statistics, in particular when the assignment-to-treatment mechanism is not controlled, as it frequently occurs in observational (causal) studies. As a matter of fact, due to the presence of confounding covariates, there could be relevant differences among subjects receiving different treatment levels. In causal studies, especially in micro-econometrics and bio-statistics, the problem of comparing a treatment vs. a control is widely studied in the literature. Bootstrap-based tests in the settings of randomized experiments are studied in [1], and permutation tests are studied in [2,3]. Nonparametric tests are proposed in [4]. The problem of comparing the effects of a treatment with several different levels, in causal studies, is much less studied. The usual approach seems to consist in focusing on pairwise ATE comparisons; cfr. [5,6]. Combinations of pairwise comparisons, for instance by Bonferroni rule, are not considered.

Consider n independent units, each receiving a treatment T with $K+1$ levels, denoted by $0, 1, \ldots, K$. The level 0 conventionally denotes the absence of treatment (control group). Let $Y_{(k)}$, $k = 0, 1, \ldots, K$ be the *potential outcome* of a subject when treatment is at level k. The *observed outcome* for a unit is then

$$Y = \sum_{k=0}^{K} Y_{(k)} I_{(T=k)} \tag{1}$$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 463–468, 2025.
https://doi.org/10.1007/978-3-031-64431-3_76

where the indicator $I_{(T=k)}$ is 1 if $T = k$, and 0 otherwise.

Treatment T has *no effect* whenever $Y_{(0)}, Y_{(1)}, \ldots, Y_{(K)}$ have the same probability distribution, *i.e.* if $Y_{(0)} \stackrel{d}{=} Y_{(1)} \stackrel{d}{=} \cdots \stackrel{d}{=} Y_{(K)}$, $\stackrel{d}{=}$ denoting equality in distribution. In the sequel, the distribution function (d.f.) of $Y_{(k)}$ is denoted by F_k, so that the treatment has no effect whenever all F_ks coincide.

The assignment-to-treatment mechanism is assumed to depend only on pre-treatment *observed* covariates, with no unobserved confounders. The relevant covariates vector is denoted by $\boldsymbol{X} = (X_1 \cdots X_P)$. The kth propensity score is the probability of receiving treatment at level k conditionally on $\boldsymbol{X} = \boldsymbol{x}$:

$$p_k(\boldsymbol{x}) = P(T = k|\boldsymbol{X} = \boldsymbol{x}); \;\; k = 0, 1, \ldots, K. \tag{2}$$

The basic assumptions on which the paper rests are listed below.

H1. *Weak unconfoundedness.* $T \perp\!\!\!\perp Y_{(k)}|\boldsymbol{X}$, for all $k = 0, 1, \ldots, K$.
H2. *Common support.* There exists a positive real ε for which $\varepsilon \leq p_k(\boldsymbol{x}) \leq 1 - \varepsilon$ for each $\boldsymbol{x}$ and $k = 0, 1, \ldots, K$.
H4. *Compactness.* The support $\mathcal{X}$ of $\boldsymbol{X}$ is a compact subset of $\mathbb{R}^P$.

Observed data are the triplets $(Y_i, T_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. The random variables (r.v.s) $(Y_i, T_i, \boldsymbol{X}_i)$ are assumed to be independent and identically distributed (*i.i.d.*).

2 Testing for No Treatment Effect

In both parametric / non- or semi-parametric situations, the focus is often on determining if any of the $K+1$ treatments does have a different effect. The main hypothesis to be tested is the *absence of treatment effect*, namely

$$\begin{cases} H_0 : \quad F_0 = F_1 = \cdots = F_K \\ H_1 : \text{The d.f.s } F_k\text{s do not coincide} \end{cases} \tag{3}$$

In order to estimate the d.f.s F_ks, an important step consists in estimating the propensity scores $p_k(\boldsymbol{x})$, $k = 0, 1, \ldots, K$. The estimator considered here, denoted by $\widehat{p}_k(\boldsymbol{x})$, is essentially a sieve estimator based on an a suitable approximation of $p_k(\boldsymbol{x})$.

Next the following estimators of F_ks are considered

$$\widehat{F}_k(y) = \sum_{i=1}^{n} I_{(Y_i \leq y)} w_{i,k}, \; k = 0, 1, \ldots, K \tag{4}$$

where

$$w_{i,k} = I_{(T_i=k)} \widehat{p}_k(\boldsymbol{x}_i)^{-1} \Big/ \sum_{j=1}^{n} I_{(T_j=k)} \widehat{p}_k(\boldsymbol{x}_j)^{-1} \; .$$

Define now $W_{k,n}(y) = \sqrt{n}(\widehat{F}_k(y) - F_k(y))$ $(k = 0, 1, \ldots, K)$, and consider the empirical process

$$\boldsymbol{W}_n(y) = [W_{0,n}(y)\, W_{1,n}(y) \cdots W_{K,n}(y)]^T \tag{5}$$

Proposition 1. *The sequence of stochastic processes (5) converges weakly, as* $n \to \infty$*, to a* $(K+1)$*-variate Gaussian process*

$$\boldsymbol{W}(y) = [W_0(y)\, W_1(y) \cdots W_K(y)]^T$$

with null mean function ($E[W_k(y)] = 0$ *for all* $k = 0, 1, \ldots, K$) *and covariance kernel* $\boldsymbol{C}(y, t) = E[\boldsymbol{W}(y) \otimes \boldsymbol{W}(t)]$*. Convergence is in the set* $l^{\infty}_{K+1}(\mathbb{R})$ *of bounded functions* $\mathbb{R} \mapsto \mathbb{R}^{K+1}$ *equipped with the sup-norm.*

We are now ready to move to the construction of a nonparametric test for the hypothesis problem (3). Our starting point consists in constructing a test that is in some way inspired to the well-known Kruskal-Wallis test used in nonparametric one-way ANOVA. As a Kruskal-Wallis type test for the hypotheses problem (3), the following test-statistic is considered

$$D_n = \sum_{k=0}^{K} \frac{n_k}{n} \left\{ \left(\sqrt{n} \int_{-\infty}^{+\infty} \widehat{F}(y)\, d\widehat{F}_k(y) - \frac{1}{2} \right) \right\}^2 \tag{6}$$

where $\widehat{F}_k$s are given by (4), and

$$n_k = \sum_{i=1}^{n} I_{(T_i=k)}, \quad k = 0, 1, \ldots, K; \tag{7}$$

$$\widehat{F}(y) = \sum_{k=0}^{K} \frac{n_k}{n} \widehat{F}_k(y). \tag{8}$$

It is worth noting that the statistic D_n in (6) can be expressed in a different but equivalent form, useful for subsequent developments. Consider first

$$\theta_{jk} = \int_{-\infty}^{+\infty} F_j(y)\, dF_k(y), \; j, k = 0, 1, \ldots, K \tag{9}$$

which is the measure of the difference between the two d.f.s F_j, F_k used in the (two-sample) Wilcoxon-Mann-Whitney statistic; cfr., for instance, [4] and references therein. Since $\theta_{jk} = 1 - \theta_{kj}$, using θ_{jk} is equivalent to use θ_{kj}. Moreover, if $j = k$ then $\theta_{kk} = 1/2$, provided F_k is continuous. Of course, when $F_0 = F_1 = \cdots = F_K$, all θ_{jk} are equal to $1/2$. To find out the limiting distribution of the test-statistic D_n under H_0, define first the estimator of θ_{jk}

$$\widehat{\theta}_{jk} = \int_{-\infty}^{+\infty} \widehat{F}_j(y)\, d\widehat{F}_k(y), \quad j \neq k = 0, 1, \ldots, K. \tag{10}$$

In terms of θ_{jk}, D_n can be written as

$$D_n = \sum_{k=0}^{K} n_k \left\{ \sum_{\substack{j=0 \\ j \neq k}}^{K} \frac{n_j}{n} \left(\widehat{\theta}_{jk} - \frac{1}{2} \right) \right\}^2 \tag{11}$$

Next step consists in studying the limiting distribution of

$$\widetilde{D}_n = \sum_{k=0}^{K} \lambda_k \left\{ \sqrt{n} \sum_{\substack{j=0 \\ j \neq k}} \lambda_j \left(\hat{\theta}_{jk} - \theta_{jk} \right) \right\}^2 = \sum_{k=0}^{K} \lambda_k T_k^2$$

where

$$T_k = \sqrt{n} \sum_{j<k}^{K} \lambda_j \left(\widehat{\theta}_{jk} - \theta_{jk} \right) - \sqrt{n} \sum_{j>k}^{K} \lambda_j \left(\widehat{\theta}_{kj} - \theta_{kj} \right), \; k = 0, 1 \ldots, K, \quad (12)$$

$$\lambda_k = P(T = k), \; k = 0, 1 \ldots, K. \quad (13)$$

Proposition 2. *Under appropriate regularity conditions, as* $n \to \infty$,

$$\widetilde{D}_n \xrightarrow{d} \boldsymbol{V}^T \boldsymbol{C} \boldsymbol{V} \quad (14)$$

where $\boldsymbol{V}$ *is a* $K(K+1)/2$*-dimensional Multivariate Normal r.v., and* $\boldsymbol{C}$ *is a symmetric, non-singular matrix.*

When $F_0 = F_1 = \cdots = F_K$, all θ_{jk} are equal to $1/2$ for all $j \neq k$, so that $\widetilde{D}_n$ reduces to

$$\widetilde{D}_n = \sum_{k=0}^{K} \lambda_k \left\{ \sqrt{n} \sum_{\substack{j=0 \\ j \neq k}}^{K} \lambda_j \left(\widehat{\theta}_{jk} - \frac{1}{2} \right) \right\}^2 \quad (15)$$

Proposition 3. *Under appropriate regularity conditions, the following two statements hold.*

1. *If* $F_0 = F_1 = \cdots = F_K$*, then*

$$D_n \xrightarrow{d} \boldsymbol{V}^T \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A} \boldsymbol{V} \;\; \text{as } n \to \infty. \quad (16)$$

   where $\boldsymbol{V}$ *is the Multivariate Normal r.v. defined in Proposition 2,* $\boldsymbol{\Lambda}$ *is the diagonal matrix of* $\lambda_0, \ldots, \Lambda_K$*, and* $\boldsymbol{A}$ *is a non-singular matrix.*
2. *If the hypothesis* $F_0 = F_1 = \cdots = F_K$ *is false, then*

$$\lim_{n \to \infty} P\left(D_n > M\right) = 1 \;\; \forall \, M > 0. \quad (17)$$

The basic idea to test for the hypothesis of absence of treatment effect is to construct a simultaneous confidence region for θ_{jk}s, based on D_n. Let $\boldsymbol{\theta}_{vec}$ be the vector with components θ_{jk}, with $j = 0, 1, \ldots, K$ and $k = j+1, \ldots, K$. The goal is to construct a confidence region of level $1 - \alpha$ for $\boldsymbol{\theta}_{vec}$, based on D_n. If the region contains the vector with all components equal to $1/2$, then H_0 is accepted, otherwise it is rejected. The confidence region for $\boldsymbol{\theta}_{vec}$ takes the form

$$\{\boldsymbol{\theta}_{vec} : \; \widetilde{D}_n \leq d_{1-\alpha}\} \quad (18)$$

where $d_{1-\alpha}$ is the $1-\alpha$th quantile of the distribution of $\widetilde{D}_n$. It can be seen that, as n gets large, the confidence region (18) contains the vector with all components equal to 1/2 if and only if $D_n \leq d_{1-\alpha}$, and hence the rejection region of the test has the form $D_n > d_{1-\alpha}$.

To approximate $d_{1-\alpha}$, a sub-sampling scheme has been developed. It consists in randomly selecting M independent sub-samples of size $m \leq n$ from the sample of Y_i, T_i, $\boldsymbol{X}_i$)s, in computing the corresponding statistic $\widetilde{D}_{m;l}$, $l = 1, \ldots, , M$, and in constructing their empirical d.f.

$$\widehat{R}_{n,m}(z) = \frac{1}{M} \sum_{l=1}^{M} I_{(\widetilde{D}_{m;l} \leq z)}.$$

and the corresponding quantile

$$\widehat{d}_{1-\alpha} = \widehat{R}^{-1}_{n,m}(1-\alpha) = \inf\{z : \ \widehat{R}_{n,m}(z) \geq 1-\alpha\}.$$

If $m \to \infty$, $m/n \to 0$ as $n \to \infty$, it is easy to show that $\widehat{d}_{1-\alpha}$ tends in probability to $d_{1-\alpha}$, for every fixed $0 < \alpha < 1$.

As a consequence, the test of no treatment effect with rejection region

$$D_n > \widehat{d}_{1-\alpha} \tag{19}$$

has asymptotic significance level α.

2.1 Testing for Stochastic Dominance

In evaluating the effect of multi-treatment, it could be of interest to test whether there is an effect on the whole distribution function of Y, namely whether changing the treatment improves the behavior of the whole d.f. of Y. This implies considering some appropriate notion of stochastic dominance for treatments effects. We consider here *first order stochastic dominance*; crf. [7]. The d.f. F_k first-order stochastically dominates F_j if $F_k(y) \leq F_j(y)$ for all real y; in this case, the notation $Y_{(j)} \preceq Y_{(k)}$ will be used.

A nice feature of the measure θ_{jk} (9) is that $Y_{(j)} \preceq Y_{(k)} \Rightarrow \theta_{jk} \geq 1/2$ (the opposite is not true), *i.e.* θ_{jk} is *monotonic* w.r.t. first order stochastic dominance. This allows its use as a test-statistic for stochastic dominance.

In the case $K = 1$ (treatment *vs.* control) different tests for stochastic dominance have been proposed in the literature; cfr. [8,9], and references therein: in [4] a Wilcoxon-type test for stochastic dominance is studied.

The goal of this section is to construct a test for the hypothesis that a certain treatment level is the "best" one, *i.e.* it stochastically dominates all other treatments. With no loss of generality it may be assumed that the treatment level to be tested as the best is K, so that the corresponding hypothesis problem is

$$\begin{cases} H_0 : & Y_{(k)} \preceq Y_{(K)} \ \forall\, k = 0, 1, \ldots, K-1 \\ H_1 : Y_{(K)} & \text{does not dominate all other treatment effects} \end{cases} \tag{20}$$

Due to the monotonicity of θ_{jk}s, the hypothesis problem (20) can be transformed in its weaker version

$$\begin{cases} H_0 : \ \theta_{k,K} \geq \frac{1}{2} & \forall\, k = 0, \ldots K-1 \\ H_1 : \theta_{k-1,k} < \frac{1}{2} & \text{for some } k = 0, \ldots K-1 \end{cases} \tag{21}$$

A simple test for the hypothesis (21) at a significance level α consists in constructing simultaneous confidence intervals for $\theta_{0K}, \theta_{1K}, \ldots, \theta_{K-1\,K}$ with overall level $1-\alpha$; H_0 is rejected if at least one of those intervals contains $1/2$.

The joint limiting distribution of $\widehat{\theta}_{jk}$ $(j < k)$, properly rescaled, is Multinormal, and corresponds to that of the r.v. $\boldsymbol{V}$ in Proposition 2. Hence, if V_{jk} $(j < k)$ denotes the generic component of $\boldsymbol{V}$, there exists a unique $d_{1-\alpha}$ for which

$$P\left(\max_{0\leq k\leq K-1} V_{k\,K} \leq d_{1-\alpha}\right) = 1-\alpha.$$

As n gets large, the K regions

$$\left[\widehat{\theta}_{k\,K} - \frac{d_{1-\alpha}}{\sqrt{n}},\, 1\right], \quad k = 0,\, 1,\, \ldots,\, K-1 \tag{22}$$

are confidence regions with approximate overall confidence level $1-\alpha$.

The quantile $d_{1-\alpha}$ can be approximated through a sub-sampling procedure similar to that of the previous Section. As a consequence, at a significance level α, the test that rejects H_0 whenever

$$\widehat{\theta}_{k\,K} - \frac{d_{1-\alpha}}{\sqrt{n}} < \frac{1}{2} \text{ for at least one } k = 0,\, 1,\, \ldots,\, K-1$$

has approximate significance level α.

References

1. Abadie, A.: Bootstrap tests for distributional treatment effects in instrumental variable models. J. Am. Stat. Assoc. **97**, 284–291 (2002)
2. Ding, P.: A paradox from randomization-based causal inference. Stat. Sci. **97**, 331–345 (2017)
3. Wu, J., Ding, P.: Randomization tests for weak null hypotheses in randomized experiments. J. Am. Stat. Assoc. **116**, 1898–1913 (2021)
4. Conti, P.L., De Giovanni, L.: Testing for the presence of treatment effect under selection on observables. AStA **107**, 641–669 (2023)
5. Yang, S., Imbens, G.W., Cui, Z., Faries, D.E., Kadziona, Z.: Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics **72**, 1055–1065 (2016)
6. Li, F., Li, F.: Propensity score matching and subclassification in observational studies with multi-level treatments. Ann. Appl. Stat. **13**, 2389–2415 (2019)
7. Anderson, G.: Nonparametric tests of stochastic dominance in income distribution. Econometrica **64**, 1183–1193 (1996)
8. Donald, S.G., Hsu, Y.C.: Estimation and inference for distribution functions and quantile functions in treatment effect models. J. Economet. **178**, 383–397 (2014)
9. Donald, S.G., Hsu, Y.C.: Improving the power of tests of stochastic dominance. Econom. Rev. **35**, 553–585 (2016)

A Bayesian Overlapping Stochastic Block Model for Clustering Biographical Networks

Noemi Corsini[1(✉)] and Michael Fop[2]

[1] Dipartimento di Scienze Statistiche, Universitá di Padova, Padova, Italy
noemi.corsini@phd.unipd.it
[2] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland
michael.fop@ucd.ie

Abstract. In the field of network analysis, community detection plays a pivotal role. As real-world networks become increasingly complex, assuming disjoint clusters often proves too restrictive. For this reason, it becomes essential to consider more flexible models that allow for overlapping clusters. The classical overlapping stochastic block model faces limitations in explicitly controlling the degree of overlap in its results. This lack of flexibility can lead to suboptimal model fit in specific scenarios. To address this challenge, we propose an extension to the overlapping stochastic block model. Our aim is to define a model that simultaneously captures and controls for both the degree of cluster overlapping and the individual centrality of each node. Unlike the estimation procedure adopted in the classical overlapping stochastic block model, we develop a Bayesian Markov chain Monte Carlo algorithm to estimate this extended version. This work finds practical application in the analysis of the Pantheon relational network. A network derived from a biographical database of the globally famous individuals throughout human history.

Keywords: network data · overlapping clusters · Bayesian estimation

1 Introduction

In the last twenty years, networks have become a fundamental tool to describe and analyse various systems, ranging from ecology to sociological structures, passing through biology and economy. For this reason, their study has also gained significant attention from a statistical perspective. One of the key features observed in many networks is the presence of communities. This refers to the phenomenon in which nodes within a network tend to form groups based on shared behaviours or characteristics. Several methods have been proposed in the statistical literature to detect these communities, including latent space models [10] and stochastic block modelling [8]. Stochastic block models (SBMs) are a class of models originally introduced by [6] in a supervised context and then extended by [4] in order to account for a latent structure of the blocks. Here, nodes are

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 469–474, 2025.
https://doi.org/10.1007/978-3-031-64431-3_77

grouped into stochastically equivalent blocks, meaning that nodes within the same community have the same probability of forming a link with another node in the network. This class of models has been thoroughly studied and extended in the literature, leading to a multitude of tailored variants, each developed for specific applications and capable to model new and ad-hoc features of the network [5]. For instance, [3] [12] proposed an extension to model the nodes' popularity, as determined by the number of edges between a node and a specific community, otherwise assumed to be equal within the same community; whereas [13,15] introduced covariates in the SBM so that each link probability would be influenced by the specific features of a node.

The standard SBM enables the detection of exhaustive and disjoint clusters. While in some scenarios this can be considered a realistic assumption, in other situations it might be quite restrictive. Therefore, methods to handle overlapping communities have been introduced. Among others, the two most known approaches are the mixed membership stochastic block model (MMSBM) developed by [11] and the overlapping stochastic block model (OSBM) proposed by [1]. In the former, nodes' cluster membership changes based on the block the node is interacting with; in the latter, nodes can belong to multiple clusters simultaneously. However, the OSBM, as currently defined, has two main limitations that can affect the flexibility of the model in certain scenarios. On one hand, it does not allow to control for the degree of cluster overlapping, and on the other hand, does not allow direct modeling of a node's centrality.

The purpose of this paper is to address these limitations and present a novel more flexible overlapping stochastic block model. The development of this extended model is justified by the need to simultaneously capture the degree of cluster overlapping and model in a flexible manner the centrality of a node with respect to the communities present in the network. The extension is motivated by the analysis of the Pantheon network data, a large biographical network of historically relevant people [16] [17]. Another important contribution of this work is the definition of the model in a Bayesian framework, which will allow uncertainty quantification.

2 The Model

A network is a collection of interconnected observations denoted as nodes, linked by edges. It generally consists of pairwise measurements, typically indicating the presence/absence of a link between pairs of nodes. These are represented through a $n \times n$ adjacency matrix denoted as $\mathbf{Y}$, where n is the total number of nodes. Each entry y_{ij} represent the connection between nodes i and j. In particular, we consider a binary network such as $y_{ij} \in \{0, 1\}$ where 1 means that there exists an edge that goes from i to j. Self-loops are excluded, resulting in the omission of diagonal elements in the adjacency matrix Y.

Consider a directed binary graph, the SBM associates to each vertex a latent variable z_i drawn from a multinomial distribution with probabilities $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, where K is the number of clusters. While the SBM assumes

that $\sum_{k=1}^{K} \alpha_k = 1$ and $\sum_{k=1}^{K} z_{ik} = 1$, $\forall i \in i = 1, \ldots, n$, the OSBM relaxes these assumptions to allow for overlapping clusters and nodes simultaneously assigned to multiple clusters. Moreover, instead of a binary cluster membership variable z_i, the OSBM introduces a K-dimensional latent vector $\mathbf{z}_i$ with binary entries indicating whether the node belongs to the corresponding cluster. Specifically,

$$\mathbf{z}_i \sim \prod_{k=1}^{K} \text{Bernoulli}(\alpha_k). \tag{1}$$

This definition allows the observations to either belong to one cluster, to multiple clusters, or to no cluster at all, allowing the identification of outlying vertices that do not fit into any specific cluster.

The standard OSBM lacks a direct mechanism to control the degree of overlapping among clusters. Specifically, it cannot be specified in a way that yields densely populated clusters with minimal overlap. The higher the value of α_k, the more populated the k-th cluster will be, and the greater the overlap tends to be with other clusters. The degree of overlap between clusters is therefore influenced by the dimension of each individual cluster. This intrinsic relationship can limit the flexibility of the model in certain scenarios. For instance, in the considered Pantheon data application, it may be desirable to spot the most central nodes, and therefore obtaining large clusters with only a few, important observations that overlap among them. This cannot be modelled by the classical OSBM which lacks the ability to explicitly control the degree of overlap in the results.

To address this limitation, the random variable o_i is introduced for each observation i. The variable takes values in the set $\mathcal{K} = \{0, \ldots, K\}$ and it quantifies the number of clusters to which node i belongs. We assume that it is distributed as a binomial distribution:

$$\begin{aligned} o_i \mid \pi_i &\sim \text{Binomial}(K; \pi_i), \\ \pi_i &\sim \text{Beta}(a_\pi, b_\pi) \end{aligned}$$

where π_i represents the propensity of node i to belong to multiple clusters simultaneously. The parameter π_i represent a node's centrality with respect to the clusters. The higher the value of π_i, the more clusters the node will belong to. Hence, the more central the node will be with respect to the clustering structure and to the network itself, as also being more likely to be connected to the other nodes. Conditioning on this quantity, the new random variable o_i allows to control for the number of overlapping clusters for a node. Given o_i, the observation is assigned to a fixed number of clusters, such that $\sum_k z_{ik} = o_i$. Consequently, to fulfill this restriction, we assume a conditional Bernoulli distribution [9] [7] as a prior on $\mathbf{z}_i$ conditional on o_i, that is $\mathbf{z}_i \mid \boldsymbol{\alpha}, o_i \sim \text{CB}(o_i, \boldsymbol{\alpha})$. Assuming that the random vectors n are all independent a priori, the joint probability mass

function of the $n \times K$ matrix $\boldsymbol{Z}$ is specified as follows

$$\mathbb{P}(\mathbf{Z} \mid \boldsymbol{\alpha}, \mathbf{o}) = \prod_{i=1}^{n} \left\{ \prod_{c=0}^{K} \left[\frac{\prod_{k=1}^{K} \left(\frac{\alpha_k}{1-\alpha_k} \right)^{z_{ik}}}{\sum_{s \in \mathcal{Z}^o} \left[\prod_{k=1}^{K} \left(\frac{\alpha_k}{1-\alpha_k} \right)^{s_k} \right]} \right]^{I(o_i=c)} \right\},$$

where $\mathcal{Z}^o$ represents the set of all valid configurations of $\boldsymbol{z}_i$ such that $\sum_{k=1}^{K} z_{ik} = o_i$ and $\boldsymbol{\alpha}$ corresponds to a K-dimensional vector representing the probabilities of belonging to each cluster, such that $\alpha_k > 0$, but it is not required that $\sum_k \alpha_k = 1$. We assume that $\alpha_k \sim \text{Beta}(a_\alpha, b_\alpha)$, $\forall k$.

Given the latent vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ for node i and j, respectively, we model the conditional probability of observing an edge between these nodes through a Bernoulli distribution

$$\mathbb{P}(y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j, \mathbf{W}) = \eta_{ij}^{y_{ij}} (1 - \eta_{ij})^{1-y_{ij}}, \tag{2}$$

where $\eta_{ij} = \Phi\left(w_0 + \mathbf{z}_i \mathbf{W} \mathbf{z}_j^\top\right)$ is the probit function. Here, $\mathbf{W}$ is a $K \times K$ matrix that contains real-valued parameters apted at determining the connectivity of the nodes within and between clusters. The intercept w_0 determines the connectivity level of nodes not assigned to any cluster, hence the density of the network if no clustering structure was present. We assume that all the entries w_{kh} for $k, h = 1, \ldots, K$ and and w_0 are independent, placing standard Normal prior distributions. Therefore, to summarize, the proposed extended overlapping stochastic block model has the following hierarchical specification:

$$\begin{aligned}
&y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j, \mathbf{W} \sim \text{Bernoulli}(\eta_{ij}), \\
&\mathbf{W} \sim \prod_{k,h} \mathcal{N}(0,1), \qquad\qquad w_0 \sim \mathcal{N}(0,1), \\
&\mathbf{z}_i \mid \boldsymbol{\alpha}, o_i \sim \text{CB}(\boldsymbol{\alpha}, o_i), \qquad \boldsymbol{\alpha} \sim \prod_{k=1}^{K} \text{Beta}(a_\alpha, b_\alpha), \\
&o_i \mid \pi_i \sim \text{Binomial}(K, \pi_i), \quad \pi_i \sim \text{Beta}(a_\pi, b_\pi).
\end{aligned}$$

3 Inference

For the proposed extended OSBM, the joint posterior distribution is given by:

$$\begin{aligned}
\mathbb{P}(\mathbf{Z}, \mathbf{W}, \mathbf{o}, \boldsymbol{\alpha}, \boldsymbol{\pi} \mid \mathbf{Y}) &\propto \mathbb{P}(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{W}, \mathbf{o}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \mathbb{P}(\mathbf{Z}, \mathbf{W}, \mathbf{o}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \\
&= \mathbb{P}(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) \mathbb{P}(\mathbf{W}) \mathbb{P}(\mathbf{Z} \mid \mathbf{o}, \boldsymbol{\alpha}) \mathbb{P}(\boldsymbol{\alpha}) \mathbb{P}(\mathbf{o} \mid \boldsymbol{\pi}) \mathbb{P}(\boldsymbol{\pi}).
\end{aligned}$$

In this expression, $\mathbb{P}(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$ denotes the likelihood as determined by eq.(2), while the remaining terms represent the prior distributions of the parameters, as described in Sect. 2.

To draw samples from the joint posterior distribution, a Markov Chain Monte Carlo (MCMC) approach is used. However, because of the model's complexity, it's impractical to find conjugacy for all parameters and sample directly from the full posterior. Consequently, an adaptive Metropolis-within-Gibbs sampler is employed. Specifically, an adaptive Metropolis-Hasting is exploited for sampling both $\mathbf{W}$ and $\boldsymbol{\alpha}$, with the adaptive step implemented following [2]. Sampling for the remaining parameters is carried out through a Gibbs sampler.

Due to the overlapping clusters, inferential challenges arise, particularly in the context of post-processing and model selection. During the post-processing step, a major difficulty lies in identifying the posterior estimate of the cluster allocation, which proves to be a challenging task. Additionally, the issue of label switching has not yet been addressed in the literature in the case of observations potentially belonging to multiple clusters.

Furthermore, since the number of clusters is often unknown, it needs to be estimated. Here we resort to model selection criteria to select the number of clusters K. Specifically, a range of models is fitted, each with a different value of K. Then, their goodness of fit is evaluated through various model selection criteria, including the Bayesian information criterion (BIC), the deviance information criterion (DIC), and the Watanabe-Akaike information criterion (WAIC).

4 Discussion and Conclusion

In this work we introduce an extended version of the overlapping stochastic block model (OSBM) that addresses some limitations of the standard OSBM. The primary motivation is to model networks with clustering structure and the presence of nodes central to the communities, with a specific focus on the Pantheon network dataset. This dataset is a biographical database of globally famous people over the last $5,500$ years. Specifically, it includes $11,341$ world-renowned subjects, who have their Wikipedia page translated in more than 25 languages. The network comprises links among these actors across their Wikipedia pages, where a link between two actors indicates that one actor is mentioned and as an hyper-link in the biography section of the other actor.

The extension of the overlapping stochastic block model (OSBM) developed in this work addresses the need to accommodate overlapping clusters while allowing to model in a flexible manner the degree of overlap. This will enable the identification of globally famous people that have a central role and connect multiple large communities, a feature that would not be possible to achieve with the standard OSBM. A key novelty is the definition of the model in a Bayesian framework, which will allow uncertainty quantification. Estimation of the OSBM is then carried out using MCMC. Compared to the variational inference framework developed for the standard OSBM, the Bayesian approach offers also additional flexibility through the specification of suitable prior distributions. However, some challenges are still open in relation to posterior inference and model selection.

References

1. Latouche, P., Birmelé, E., Ambroise, C.: Overlapping stochastic block models with application to the French political blogosphere. Ann. Appl. Stat. **5**(1), 309–336 (2011). https://doi.org/10.1214/10-AOAS382
2. Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**, 343–373 (2008)
3. Sengupta, S., Chen, Y.: A block model for node popularity in networks with community structure. J. R. Stat. Soc. Ser. B Stat. Methodol. **80**(2), 365–386 (2018)
4. Krzysztof, N., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. J. Am. Stat. Assoc. **96**(455), 1077–1087 (2001)
5. Lee, C., Wilkinson, D.J.: A review of stochastic block models and extensions for graph clustering. Appl. Netw. Sci. **4**(1), 1–50 (2019). https://doi.org/10.1007/s41109-019-0232-2
6. Wang, Y.J., Wong, G.Y.: Stochastic blockmodels for directed graphs. J. Am. Stat. Assoc. **82**(397), 8–19 (1987)
7. Chen, S.X.: General properties and estimation of conditional Bernoulli models. J. Multivar. Anal. **74**(1), 69–87 (2000)
8. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)
9. Chen, S.X., Liu, J.S.: Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. Stat. Sinica **7**(4), 875–892 (1997)
10. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. J. Am. Stat. Assoc. **97**(460), 1090–1098 (2002)
11. Airoldi, E.M., et al.: Mixed membership stochastic blockmodels. Adv. Neural Inf. Process. Syst. **21** (2008)
12. Noroozi, M., Rimal, R., Pensky, M.: Estimation and clustering in popularity adjusted block model. J. R. Stat. Soc. Ser. B Stat. Methodol. **83**(2), 293–317 (2021)
13. White, A., Murphy, T.B.: Mixed-membership of experts stochastic blockmodel. Netw. Sci. **4**(1), 48–80 (2016)
14. Handcock, M.S., Raftery, A.E., Tantrum, J.M.: Model-based clustering for social networks. J. R. Stat. Soc. A. Stat. Soc. **170**(2), 301–354 (2007)
15. Sweet, T.M.: Incorporating covariates into stochastic blockmodels. J. Educ. Behav. Stat. **40**(6), 635–664 (2015)
16. Yu, A.Z., Ronen, S., Hu, K., Lu, T., Hidalgo, C.A.: Pantheon 1.0, a manually verified dataset of globally famous biographies. Sci. Data **3**(1), 1–16 (2016)
17. Beytía, P., Schobin, J.: Networked pantheon: a relational database of globally famous people: social and behavioural sciences. Res. Data J. Humanit. Soc. Sci. **5**(1), 50–65 (2020)

Exploring Sustainability Amidst Global Challenges: A Focus on Cooperative Enterprises

Michele Costa(✉) and Flavio Delbono

Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy
{michele.costa,flavio.delbono}@unibo.it

Abstract. In this study, we address sustainability issues from the perspective of the entrepreneurial model, comparing cooperative and capitalist microenterprises based on the labor cost to value-added ratio. The analysis employs indicators such as the probability and intensity of transvariation, along with the Gini inequality index, highlighting the significant impact of cooperatives. Overall, the study underscores the pivotal role of cooperative enterprises in sustainable development and emphasizes the need to address their challenges for long-term resilience.

Keywords: Sustainability · Cooperatives · Transvariation

1 Sustainability and Cooperation

The themes of sustainability may not have been given the attention they deserve, especially amidst the challenges posed by the pandemic, conflicts in Ukraine and the Middle East, and the surge in inflation, which have shifted focus elsewhere. It's imperative to question their continued relevance in ongoing conversations and explore strategies to bring them back into the forefront of the agenda.

In our examination of sustainability, we emphasize the crucial role played by a structural feature [10], such as the entrepreneurial model, highlighting the role of cooperative enterprises as frontrunners in advancing the SDG 2030 agenda [11]. Cooperatives serve as vital agents in achieving SDGs, embodying principles of collaboration and community engagement. Moreover, cooperatives, as a form of enterprise, exhibit a notable dedication to addressing employment concerns [1, 2], demonstrating a willingness to sacrifice profits for the benefit of employment and labor remuneration. Their commitment to fostering solidarity within communities aligns closely with the principles of sustainable development. By emphasizing cooperative values such as democratic decision-making and equitable distribution of resources, these enterprises not only contribute to the achievement of SDGs but also promote decent work and economic growth (SDG 8) and reduce inequalities (SDG 10). As we delve into the comparison between cooperative and capitalist microenterprises, we scrutinize the labor cost to value-added ratio to better understand their respective contributions towards sustainable development. Our results show that this ratio is consistently higher for cooperatives, confirming their role as a sustainable entrepreneurial model, and the differences between cooperative and capitalist microenterprises appear to be increasing.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 475–479, 2025.
https://doi.org/10.1007/978-3-031-64431-3_78

2 Data and Methods

We conduct the comparison at the micro level (enterprises with fewer than 10 employees and a turnover of less than two million euros) because the micro context better highlights the specificities of cooperative enterprises as key players in the implementation of the SDGs. We utilize an individual-level database that provides a strong informational detail, allowing comparison between entire distributions of variables and not just in reference to aggregated values and synthetic moments of distribution like the mean. The period covered by our database ranges from 2012 to 2021.

We carry out the comparison between cooperative microenterprises and capitalist microenterprises based on the ratio of labor cost to value added (y). In particular, we evaluate the difference between cooperatives (I group, of size n_1) and capitalists (II group, of size n_2) based on the differences (y_{1i} - y_{2r}). Among the various indicators to analyze these differences, two measures proposed by Gini [7–9] are considered.

The first measure we consider is the probability of transvariation P. Given two groups, P refers to a situation where the median of the first group is greater than the median of the second group, but at least one difference between a unit i in first group and a unit r in second group is negative (y_{1i} - y_{2r}) < 0.

P is computed as the number of overlapping differences to its maximum, that is as:

$$P = (2 *_A n +_B n)/(n_1 * n_2) \tag{1}$$

where $_A$n and $_B$n are, respectively, the number of negative and the number of null differences (y_{1i} - y_{2r}). The maximum of P is reached when the two groups completely overlap and the median of the first group equals the median of the second group. Probability of transvariation ranges between 0, when no difference is overlapping, and 1, when overlapping component reaches its maximum.

The second indicator, the intensity of transvariation, I, refers to a situation where the mean of the first group is greater than the mean of the second group, but at least one difference between a unit in the first group and a unit in the second group is negative: (y_{1i}-y_{2r}) < 0. Overlapping in this case is defined by the relationship between the means.

With respect to P, which considers the number of overlapping differences, I is based on their size, that is on the quantities $|y_{2r} - y_{1i}|$ for $\bar{y}_1 > \bar{y}_2$, $y_{2r} > y_{1i}$.

Let T be the sum of overlapping differences, then the intensity of transvariation is obtained as the ratio of the sum of overlapping differences T to its maximum:

$$I = 2T / \sum_{r=1}^{n_2} \sum_{i=1}^{n_1} |y_{2r} - y_{1i}| \tag{2}$$

Further within the analysis of differences (y_{1i} - y_{2r}), another crucial reference is the Gini inequality index, which, in a population disaggregated into 2 groups, is

$$G = \frac{1}{2n^2\bar{y}} \sum_{j=1}^{2} \sum_{h=1}^{2} \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \tag{3}$$

Alongside G, following [5] we also introduce G_{jh}

$$G_{jh} = \frac{1}{n_j n_h (\bar{y}_j + \bar{y}_h)} \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}|. \tag{4}$$

For $j = h$, G_{jj} is the usual Gini index in the j-th group. By setting $p_j = n_j/n$ and $s_j = p_j\bar{y}_j/\bar{y}$, the Gini index can be obtained as a weighted sum of G_{jh} with weight $p_j s_h$

$$G = \sum_{j=1}^{2}\sum_{h=1}^{2} G_{jh}p_j s_h \tag{5}$$

In this framework, the inequality within *2* groups is $G_w = G_{11}p_1 s_1 + G_{22}p_2 s_2$, while the contribution to total inequality attributable to the differences between *2* groups can be measured as [3, 6, 12]

$$G - G_w = \sum_{j=1}^{2}\sum_{h=1}^{2} G_{jh}p_j s_h \quad j \neq h \tag{6}$$

that is a quantity which takes account of all possible differences between 2 groups.

Dagum suggests a further decomposition of the component $(G - G_w)$, by distinguishing the net Gini inequality between groups

$$G_b = \sum_{j=1}^{2}\sum_{h=1}^{2} G_{jh}D_{jh}p_j s_h \quad j \neq h \tag{7}$$

and the contribution of the intensity of transvariation

$$G_t = \sum_{j=1}^{2}\sum_{h=1}^{2} G_{jh}(1 - D_{jh})p_j s_h \quad j \neq h \tag{8}$$

where D_{jh} is the directional economic distance ratio [4, 5], which is defined as $D_{jh} = (d_{jh} - p_{jh})/(d_{jh} + p_{jh})$, with $\bar{y}_j > \bar{y}_h$.

The directional economic distance ratio takes value in the closed interval [0,1], with $D_{jh} = 0$ when $\bar{y}_j = \bar{y}_h$ and $D_{jh} = 1$ when the two group distributions do not overlap. The directional economic distance ratio is a function of d_{jh}, the gross economic affluence between the j-th and the h-th groups, and of p_{jh}, the first order moment of transvariation between the j-th and the h-th groups. Given $\bar{y}_j > \bar{y}_h$, the gross economic affluence between the groups j and h, d_{jh}, is [4, 5] a weighted average of the differences $(y_{ji} - y_{hr})$ for $y_{ji} > y_{hr}$ and can be expressed as $d_{jh} = E_j(yF_h(y)) + E_h(yF_j(y)) - E_h(y)$ where the subscripts indicate the groups j and h, $F(y)$ is the cumulative distribution function and E stands for the mathematical expectation operator. Furthermore, the first-order moment of transvariation between the j-th and the h-th groups, p_{jh}, is a weighted average of the differences $(y_{ji} - y_{hr})$ for all $y_{ji} < y_{hr}$; following [4, 5] p_{jh} can be expressed as $p_{jh} = E_j(yF_h(y)) + E_h\big(yF_j(y)\big) - E_j(y)$.

By jointly using the probability P and the intensity I of transvariation, the net Gini inequality between groups G_b and the contribution of the transvariation G_t, it is possible to obtain a wide information set about the relevance and the extent of the overlapping component and a throughout analysis of the differences between the groups.

3 Results and Conclusions

The dynamics observed in microenterprises from 2012 to 2021 underscore, in addition to the anticipated negative impact during the pandemic, significant challenges for cooperative enterprises. Their numbers have consistently declined every year since 2018,

whereas in preceding years, their growth was markedly more subdued compared to capitalist enterprises. However, despite these trends, the ratio of labor cost to value added remains systematically and significantly higher in cooperative. As cooperatives represent an important channel for achieving the SDGs, their decline constitutes a signal to be taken into account (Table 1).

In the first column of Table 2, the probability of transvariation, P, is presented. It demonstrates a declining trend over time, signifying a growing differentiation between the two groups and suggesting a more significant role for the cooperative system. Interestingly, in 2020, marked by the onset of COVID-19, P experiences an increase, indicating that the pandemic, impacting the entire economy, has mitigated the disparities between types of enterprises. A similar trend, with even more pronounced effects, is echoed by I, which decreases by nearly 10 points from 2012 to 2021. This underscores a heightened influence of the cooperative framework on labor costs relative to value added. These findings are corroborated by the ratios G_b/G and G_t/G. Overall, while the inequality between the two groups remains relatively stable, the transvariation component G_t undergoes a marked decline during the period under scrutiny.

The availability of individual data allows for the development of a bootstrap procedure to validate the robustness of our results. Through this procedure, companies are resampled from the original list using an i.i.d. sampling with replacement. For each time period examined, 1000 bootstrap replications are calculated. The test between the observed values reported in Table 2 and the mean of the bootstrap distribution does not indicate significant differences. This indicates that our findings are robust and lends general validity to our results.

In summary, our study underscores the pivotal role of cooperative enterprises in advancing sustainable development goals and highlights the necessity of addressing the challenges these enterprises face to ensure their long-term resilience and continuity.

Table 1. Cooperative and capitalist enterprises: number and labor cost to value added ratio.

	N coop	N cap	Y coop	Y cap
2012	20.741	287.444	0,841	0,698
2013	21.187	294.919	0,843	0,701
2014	22.279	317.526	0,835	0,682
2015	22.880	330.480	0,825	0,667
2016	22.516	335.097	0,825	0,673
2017	22.385	340.994	0,833	0,669
2018	21.572	339.606	0,840	0,667
2019	20.778	351.836	0,841	0,673
2020	18.331	335.382	0,815	0,661
2021	17.116	334.077	0,795	0,604

Table 2. Probability and intensity of transvariation, contribution to overall inequality by inequality between and transvariation component.

	P	I	G_b/G	G_t/G
2012	0,640	0,560	0,089	0,031
2013	0,641	0,562	0,091	0,029
2014	0,624	0,541	0,086	0,031
2015	0,628	0,556	0,089	0,030
2016	0,601	0,522	0,088	0,027
2017	0,592	0,504	0,086	0,024
2018	0,589	0,497	0,083	0,023
2019	0,590	0,496	0,075	0,022
2020	0,612	0,513	0,076	0,022
2021	0,571	0,473	0,073	0,018

Acknowledgments. Prin 2020 grant 2020B2AKFW is gratefully acknowledged.

References

1. Burdin, C., Dean, A.: Revisiting the objectives of worker cooperatives: an empirical assessment. Econ. Syst. **36**, 158–171 (2012)
2. Caselli, G., Costa, M., Delbono, F.: What do cooperative firms maximize, if at all? evidence from emilia-romagna in the pre-covid decade. Annals of Public and Cooperative Economics **93**, 821–847 (2022)
3. Costa, M.: Transvariation and inequality between subpopulations in the Dagum's Gini index decomposition. Metron **67**, 229–241 (2009)
4. Dagum, C.: Inequality measures between income distributions with applications. Econometrica **48**, 1791–1803 (1980)
5. Dagum, C.: Measuring the economic affluence between populations of income receivers. J. Business and Economic Statistics **5**, 5–12 (1987)
6. Dagum, C.: A new decomposition of the Gini income inequality ratio. Empirical Economics **22**, 515–531 (1997)
7. Deutsch, J., Silber, J.: Gini's "Transvariazione" and the measurement of distance between distributions. Empirical Econ. **22**, 547–554 (1997)
8. Gini, C.: Transvariazione, Libreria goliardica, Roma (1959)
9. Giorgi, G.M.: Gini's scientific work: an evergreen. Metron **63**, 299–315 (2005)
10. Jollès, M., Meyermans, E., Borek, V.: Determinants of macroeconomic resilience in the euro area: an empirical assessment of national policy layers. Economic Syst. 101093 (2003)
11. Lafont, J., Ramon, J., Ribeiro-Soriano, D.: The role of cooperatives in sustainable development goals. Resour. Policy. Policy **83**, 103670 (2023)
12. Mehran, F.: A statistical analysis of income inequality based on a decomposition of the Gini index, Bulletin of the International Statistical Institute, 40th Session, pp. 145–150 (1975)

Goodness of Fit Measures for Generalized Estimating Equations. An Empirical Evidence on Relationships Between Gender Diversity and Corporate Performance

Anna Crisci[1(✉)], Luana Serino[2], Pasquale Sarnacchiaro[1], and Francesco Campanella[3]

[1] University of Naples Federico II, Napoli, Italy
{anna.crisci,pasquale.sarnacchiaro}@unina.it
[2] Pegaso Telematic University, Napoli, Italy
luana.serino@unipegaso.it
[3] University of Campania "Luigi Vanvitelli", Caserta, Italy
francesco.campanella@unicampania.it

Abstract. The principal idea behind generalized estimating equations is to generalize and extend the usual likelihood score equation for a generalized linear model by including the covariance matrix of the clustered responses. The advantage of the GEE is that we do not need to specify the entire response distribution. Only the mean structure, the mean–variance relationship and specification of the covariance structure need to be defined. In order to evaluate the strength of association between a response variable and covariates, coefficient of determination based on Wald Statistics, Likelihood Ratio and pseudo-R^2 based on a quasi-likelihood method are presented and some relationships among them are shown. Following the strand of literature that examines the relationship between gender diversity and corporate performance, this study aims to analyze whether and how the presence of women in corporate governance bodies affects the corporate performance (ROA) of companies belonging to the European banking sector.

Keywords: R^2 measures · Wald statistics · Likelihood Ratio statistics · gender diversity · corporate social responsibility

1 Introduction

Generalized Estimating Equation (GEE) [7, 11] is a general statistical approach to fit a marginal model for longitudinal data analysis. GEEs have become an important strategy in the analysis of correlated data. These data sets can arise from longitudinal studies, in which subjects are measured at different points in time, in which measurements are taken on subjects who share a common characteristic, such as belonging to the same litter. GEE is a population-based approach based on a quasi-likelihood function, and provides the marginal or "population averaged" estimate of the parameters. The main idea behind

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 480–484, 2025.
https://doi.org/10.1007/978-3-031-64431-3_79

GEE is to generalize and extend the usual likelihood equation for a Generalized Linear Model by including the covariance matrix of the responses. The biggest advantage of the GEE is that we do not need to specify the entire response distribution. Only the mean structure, the mean–variance relationship and specification of the covariance structure need to be defined.

1.1 Notation

Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{it_i})\prime$ be a vector of response values, and let $\boldsymbol{X_i} = (\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{it_i}})'$ be a $t_i \times K$ matrix of covariates, with $\boldsymbol{x_{it}} = (x_{it1}, \ldots, x_{itK})\prime$, i = 1,2, …,n, where n is the number of clusters and t = 1,2, …, t_i. To simplify notation, let $t_i = T$ without loss of generality. The expected value and variance of measurement y_{it} can be expressed using a generalized linear model: $E(y_{it}|\boldsymbol{x_{it}}) = \mu_{it}$. Assume that the regression model is $\eta_{it} = g(\mu_{it}) = \boldsymbol{x_{it}}\prime\boldsymbol{\beta}$ where g is a link function and $\boldsymbol{\beta}$ is an unknown $K \times 1$ vector of regression coefficients. The $Var(y_{it}|\boldsymbol{x_{it}}) = \upsilon(\mu_{it})\phi$, where υ is a variance function of μ_{it} and ϕ is a scale parameter which may need to be estimated. Typically, υ and ϕ depend on the outcome distributions. The variance-covariance matrix for $\boldsymbol{y_i}$ is noted by $\boldsymbol{V}_i = \phi \boldsymbol{A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}}$, $\boldsymbol{A}_i = diag\{\upsilon(\mu_{i1}), \ldots, \upsilon(\mu_{iT})\}$ and the so-called "working" correlation structure $\boldsymbol{R_i(\alpha)}$ describes the pattern of measures within the subjects, which is of size $T \times T$ and depends on a vector of association parameters denoted by $\boldsymbol{\alpha}$. The GEE estimate of $\boldsymbol{\beta}$ is obtained by solving: $U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i\prime[V(\widehat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{s_i} = 0$ where $\boldsymbol{s}_i = \left(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i\right)$ with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} = (\mu_1, \ldots, \ldots \mu_{iT})\prime$ and $\left(\widehat{\boldsymbol{\alpha}}\right)$ is a consistent estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{D}_i\prime = \boldsymbol{X}_i\prime\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_i = diag(\partial_{\boldsymbol{\mu}_{i1}}/\partial_{\boldsymbol{\eta}_{i1}} \ldots\ldots., \partial_{\boldsymbol{\mu}_{it}}/\partial_{\boldsymbol{\eta}_{it}})$. Under certain regularity conditions $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed [7] with a mean $\boldsymbol{\beta}_0$ and covariance matrix estimated based on naïve or model based (a) or robust, empirical or sandwich estimator (b):

$$V_{model}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{m} D_i' V_i^{-1} D_i\right)^{-1} = M^{-1}(a)$$

$$\boldsymbol{V_{empirical}}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{M}^{-1}\left(\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{s_i} \boldsymbol{s_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i}\right) \boldsymbol{M}^{-1}(\mathbf{b})$$

2 $\widetilde{\boldsymbol{R}}^{\boldsymbol{2}}_{\widetilde{\boldsymbol{W}}}$ and $\widetilde{\boldsymbol{R}}^{\boldsymbol{2}}_{\boldsymbol{LR}}$ for GEE Model

In this section, we consider the coefficient of determination in GEE based on Wald statistics [5, 6, 8, 9]:

$$\tilde{R}^2_{\tilde{W}} = \frac{\tilde{W}}{n + \tilde{W}} \tag{1}$$

where

$$\tilde{W} = \left[\boldsymbol{C}\widehat{\boldsymbol{\beta}}\right]\prime\left[\boldsymbol{C}\widehat{\boldsymbol{V}}^{\boldsymbol{R}}\boldsymbol{C}\prime\right]^{-1}\left[\boldsymbol{C}\widehat{\boldsymbol{\beta}}\right]$$

is the Wald statistics with the GEE robust covariance matrix estimated under the null and denoted by $\widehat{\boldsymbol{V}}^{\boldsymbol{R}}$ and $\boldsymbol{C}$ is a $(K-1)\times K$ matrix with its first column having all 0s, and its last $(K-1)$ columns being the $(K-1)$ identity matrix.

Buse (1973) provided an R^2 suitable for the generalized linear regression model. His R^2 bears the same relation to the appropriate F statistic as does the standard R^2 in the basic model. Buse' s R^2 is an R^2_W.

Later, we consider the LR statistics related to Wald statistic with the GEE robust covariance matrix:

$$\tilde{LR} = n\log\left(1+\frac{\tilde{W}}{n}\right) \tag{2}$$

From (1) and (2), we obtained $\tilde{R}^2_{LR}$

$$\tilde{R}^2_{LR} = 1-\exp\left(-\frac{\tilde{LR}}{n}\right)$$

Finally, in order to generalize the pseudo- $R^2_{McFadden}$ (1974) to GEE, the maximum likelihood calculations must be replaced by quasi-likelihood ones. The pseudo-$R^2_{McFadden}$ based on quasi-likelihood is:

$$R^2_{McFadden} = 1-\frac{Q(M_\beta)}{Q(M_\alpha)}$$

where $Q(M_\alpha)$ is the quasi-likelihood for the model with only an intercept, and (M_β) is the model with intercept and independent variables.

3 Results

Following the strand of literature that examines the relationship between gender diversity and corporate performance, this study aims to analyze whether and how the presence of women in corporate governance bodies affects the corporate performance (ROA) of companies belonging to the European banking sector. The empirical analysis was carried out on a sample of 1874 European listed banks over the period 2019–2022. The final sample consists of 7496 bank-year observations. Furthermore, the data of this paper is collected from two data sources: data on gender were obtained from the Refinitiv Eikon database, formerly called "Thomson Reuters", while bank financial variables were sourced from the Bankfocus database. ESG data is obtained from the Thomson Reuters Datastream database under the category ESG – Asset 4.5. The overall ESG score is an aggregated measure of the social performance in several environmental, social and governmental categories (e.g. Employment Quality, Health & Safety, Training & Development, Human Rights, Community).

As shown in Table 1, the variables are all significant except for the variables Debt-to-Equity ratio and Independent Directors (p. value >0.05). More specifically, all the variables associated with women's presence, encompassing women employees, women in top management, and women in the Board of European Banks have positive and statistically significant coefficients, revealing that the inclusion of women in a bank's human

capital is favorably associated with better financial performance [1, 10]. Moreover, the variable Board Background exhibits a positive and statistically significant coefficient. This underscores the crucial role of board expertise for achieving positive enterprise performance [4]. Hence, our study emphasizes the pivotal contribution of BoD's member experience, especially in the banking sector, towards attaining favorable financial outcomes. Lastly, regarding board size, the coefficient is positive and statistically significant. This implies that the size of the board in a European bank has a positive impact on corporate profitability. The larger board size is associated with increased diversity in terms of stakeholder representation, enhancing efficiency in handling responsibilities [2]. In order to evaluate the goodness of fit for GEE model, we obtained $\tilde{R}^2_{\tilde{W}} = 0.40$ and $R^2_{McFadden} = 0.34$. We, also, replace the Wald statistics with the GEE robust covariance matrix in Langrage Multiplier (LM) statistic.

Furthermore, we have the following relationships: $4920.37\left(\tilde{W}\right) \geq 3782.82(LR) \geq 2970.52(LM)$.

Table 1. Parameter estimates for GEE model with normal regression model and exchangeable working correlation matrix.

Coefficients	Estimate	Robust S.E	Z	p.value
Debt-to-equity ratio	−2.819	1.764	−1.60	0.110
Independent Directors	0.165	0.857	0.19	0.847
Board Background	6.999	0.564	12.40	0.000**
Board Size	2.227	0.850	2.62	0.009**
Women directors	0.246	0.009	26.30	0.000**
Women employees	0.172	0.010	16.90	0.000**
Women managers	0.214	0.009	23.16	0.000**
_cons	8.767	1.111	7.89	0.000**

$\tilde{W} = 4920.37$; LR = 3782.82; LM = 2970.52; Prob > Chi-square 0.000; **significant at 5%

References

1. Arzu, D., Mantovani, G.M.: The Gender Contribution to the Corporate Governance and the Corporate Performance. SSRN 2885413 (2016)
2. Birindelli, G., Dell'Atti, S., Iannuzzi, A.P., Savioli, M.: Composition and activity of the board of directors: Impact on ESG performance in the banking system. Sustainability **10**, 4699 (2018)
3. Buse, A.: Goodness of fit in generalized least squares estimation. Am. Stat. **27**, 106–108 (1973)
4. Carter, C., Lorsch, J.: Back to the Drawing Board Designing Corporate Boards for a Complex World. Harvard Business School Press, Boston, MA (2004)
5. Crisci, A.: A note on regression diagnostics for generalized estimating equations: empirical study on environmental disclosure determinants. J. Operational Res. Society **74**(4), 1042–1048 (2023)

6. Crisci, A., D'Ambra, L., Esposito,V.: A generalized estimating equation in longitudinal data to determine an efficiency indicator for football teams. Social Indicator Research, Springer **146**(1–2), p. 249–261, 1573–0921 (2019)
7. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**(1), 13–22 (1986)
8. Magee, L.: R2 measures based on wald and likelihood ratio joint significance tests. Am. Stat. **44**(3), 250–253 (1990)
9. Natarajan, S., Lipsitz, S., Parzen, M., Lipshultz, S.: A measure of partial association for generalize estimating equations. Stat. Model. **7**(2), 175–190 (2007)
10. Serino, L., Papa, A., Campanella, F., Mueller, J.: Credit access and performance of Italian firms: how relevant is gender? Int. J. Manag. Financ. Account. **11**(3–4), 269–289 (2019)
11. Zeger, S.L., Liang, K.Y.: Longitudinal data analysis for discrete and continuous outcomes. Biometrics **42**(1), 121–130 (1986)

Educational Disparities in the Risk of Dementia and Subsequent Risk of Hospitalization in Lazio Region

Agostino Cristofalo[1](✉), Silvia Cascini[2], Giulia Cesaroni[2], Eleonora Trappolini[1], Nera Agabiti[2], and Anna Maria Bargagli[2]

[1] Sapienza University of Rome, Rome, Italy
agostino.cristofalo@uniroma1.it
[2] Department of Epidemiology – Lazio Regional Health Service, Rome, Italy

Abstract. Against the backdrop of population aging, dementia is an increasingly relevant issue for population health. Education plays a key role in this regard, influencing the risk of developing dementia and serving as a crucial dimension in health disparities among older adults. While many studies indicate a protective effect of more years of education on the risk of dementia, the extent to which education shapes inequalities beyond the onset of the disease remains underexplored. In this study, we analyze educational differences in the risk of developing disease and in healthcare utilization among people living with dementia. Regarding the latter, we focus on the risk of first hospitalization after dementia identification. The analysis is conducted in the Lazio Region (Italy) between 2012 and 2022. We use the Lazio Longitudinal Study, combining 2011 Census data (resident population aged 50–90) and information from regional healthcare administrative databases through record linkage. We estimate Kaplan-Meier survival curves by level of education for dementia identification among disease-free individuals, and cumulative hazard functions for the risk of first subsequent hospitalization among patients identified as incident cases. Combining population-comprehensive data sources, our study contributes to current knowledge by offering novel evidence on the educational gradient in dementia incidence and healthcare utilization in Italy.

Keywords: Health disparities · dementia · aging population · Italy

1 Background

Against the backdrop of population aging, dementia has become an increasingly relevant issue for population health. As the risk of dementia increases with age, the number of dementia cases worldwide is expected to triple by 2050 due to increasing life expectancy [1]. Although the rise of dementia is inevitable in these circumstances, recent evidence suggests that dementia incidence is declining at least in high-income countries [2]. Such a decline has been driven by general improvements in lifestyles, cardiovascular health, and increased educational opportunities [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 485–490, 2025.
https://doi.org/10.1007/978-3-031-64431-3_80

Education is considered a protective factor against dementia for several reasons. First, it provides greater cognitive reserve, which delays the cognitive decline despite the beginning of brain deterioration [3, 4]. Education is also associated with healthier behaviors and more frequent access to preventive healthcare [5, 6]. Lastly, individuals with lower education are more exposed to chronic stress caused by insecure life circumstances, isolation, and discrimination [7], resulting in higher dementia risk [8].

Education is also an important dimension of health disparity in society, especially at older ages, and it might contribute to disparities far beyond the dementia onset (cit). While differences in the risk of developing dementia by educational attainment are well-documented [4, 7, 9], less is known about the role of education once dementia has already occurred, especially as far as healthcare utilization is concerned. Some studies show that dementia patients use services more extensively, especially for preventable causes [10, 11]. Higher educated individuals generally access specialist services more and have lower avoidable mortality rates [6, 12]. To the best of our knowledge, no recent studies have focused specifically on the association between education and healthcare utilization among individuals with dementia.

2 Aim of the Study and Contribution

This study aims shed light on educational disparities in the risk of developing dementia and subsequent hospitalization among people living with in the Lazio Region (Italy) from 2012 to 2022.

The contribution of this study is twofold. First, the study uses valuable population-comprehensive health records to evaluate educational disparities in dementia risk. This is pioneering in the Italian context, since previous evidence on educational disparities in dementia incidence in Italy is based on prospective studies in small areas [13, 14]. Second, it explores educational disparities beyond dementia onset, particularly focusing on the first hospitalization after dementia identification among incident cases.

3 Data and Methods

3.1 Data

We use the Lazio Region Longitudinal Study, which combines the 2011 Census cohort of residents in the region with health records from regional administrative databases, through record linkage procedures. More precisely, we use the following administrative data sources: the Hospital Discharge Registry (HDR), the Drug Claims Registry (PHARM), and the Ticket Exemption Registry (TER).

3.2 Research Design

We carry out a retrospective cohort study. In the first part of the analysis, we enroll all residents in Lazio aged 50–90 who completed the 2011 Census questionnaire and were not identified as dementia cases (i.e., dementia-free) by 1st January 2012 (N = 916,137

men; N = 1,099,385 women). For dementia identification, we employ a validated algorithm [15] which considers individuals as dementia cases if they meet at least one of the following conditions: a) at least two different prescriptions of drugs for dementia within twelve months in the PHARM; b) at least one hospital discharge with a primary or secondary diagnosis of dementia in the HDR; c) reported exemption from healthcare co-payment specific to dementia in the TER. To ascertain whether individuals are dementia-free, we use a five years look-back period before the beginning of the observational window. We then follow dementia-free individuals are identified with dementia, censored, or until the end of follow-up (31st December 2022).

In the second part of the analysis, we enroll the incident cases retrieved from the first stage of analysis (N = 24,551 men; N = 34,454 women) and follow them until the first hospitalization after dementia identification, censoring, or the end of follow-up (31st December 2022). Hospitalizations are retrieved from the HDR.

3.3 Methods

To achieve our aim, we conduct two separate analyses. In the first analysis, we estimate Kaplan-Meier survival curves for dementia identification. Survival is considered from the beginning of follow-up, i.e. on 1st January 2012, until the earliest fulfillment of the conditions denoting dementia identification. In the second analysis, we estimate cumulative hazard functions for subsequent hospitalization. Here, individuals are considered exposed to risk from the time they are identified with dementia until the first hospitalization, if any. In both analyses, the exposure variable is education, which is retrieved from the Census and is categorized into three levels: up to lower-secondary (low), upper-secondary (middle), and tertiary (high). Individuals are right-censored if they do not appear in health records for more than twelve months, at age 90 (included), at death, and at the end of the observational period on 31st December 2022. We carry out the analyses separately for men and women, using continuous age as time scale.

4 Results

4.1 Dementia Identification

We observe a striking educational gradient in the risk of dementia identification for both men and women (see Fig. 1). Consistently across sexes, the lower educated show a remarkably higher risk of being identified with dementia, while the opposite holds true for the higher educated. The risk of the middle educated is slightly different between men and women. Whereas middle educated men show a significant divergence from the higher educated at the oldest age, middle educated women show a more similar survival to those with higher education at all ages.

4.2 Subsequent Hospitalization

Regarding the first hospitalization after dementia identification, the data present a more complex picture and more heterogeneity for men and women (see Fig. 2). On one hand,

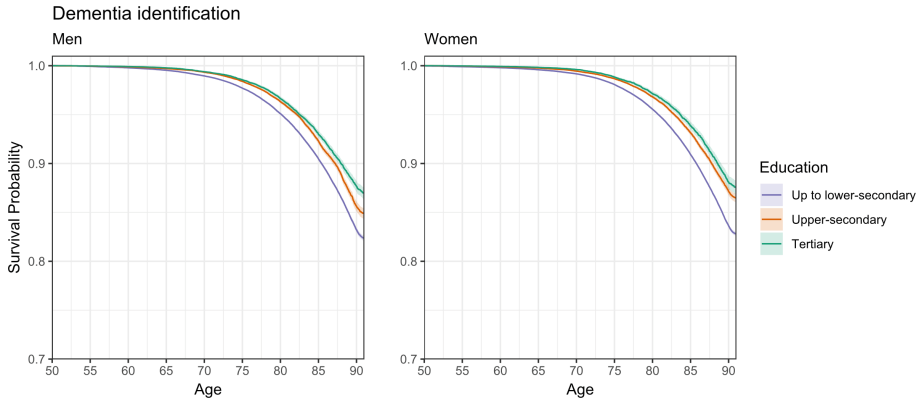


Fig. 1. Kaplan-Meier survival curves for dementia identification by level of education in Lazio among men (left-hand side) and women (right-hand side). Lazio Longitudinal Study (2012–2022). Authors' elaboration. Note: Surrounding areas are 95% C.I.

men clearly exhibit an educational gradient, with the lower educated showing a much higher cumulative hazard compared to the middle and higher educated from age 50. On the other hand, women's cumulative hazard is more similar across educational groups. At the earliest ages considered, middle educated women have higher risk compared to lower and higher educated. Around age 65, the cumulative hazard of the lower educated converges with the middle educated, whereas the hazard for higher educated women remains slightly lower. However, the differences between groups do not reach significance at 95%.

5 Conclusion

This study uses population comprehensive health records to investigate educational disparities in the risk of developing dementia and healthcare utilization after disease onset. We are able to identify an educational gradient in the risk of dementia onset, consistently with previous research [4, 9]. For both men and women, low education is confirmed as an important risk factor for dementia.

For what concerns the educational disparities in healthcare utilization after dementia identification – particularly in the first subsequent hospitalization – we found a clear educational gradient among men and less heterogeneity among women. This result suggests a gendered pattern in educational disparities in healthcare utilization among dementia patients.

The next steps of this research consist in modelling the risk of dementia onset and subsequent healthcare utilization using proportional hazard regression models and several outcomes of healthcare utilization.

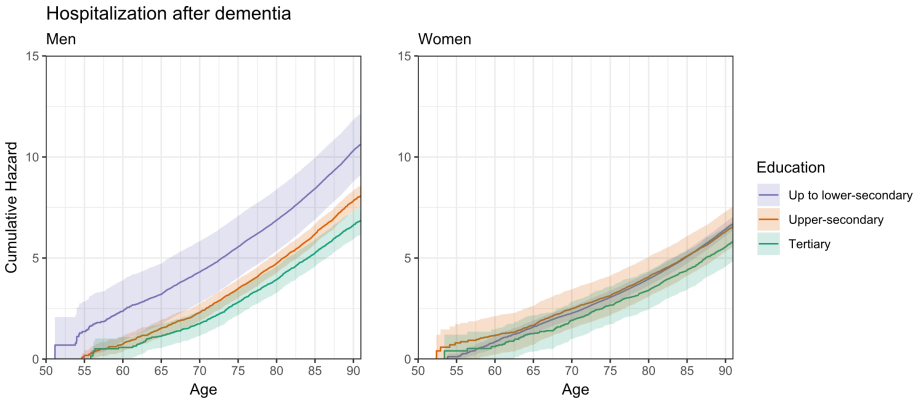


Fig. 2. Cumulative hazard functions for first hospitalization after dementia identification among men (left-hand side) and women (right-hand side). Lazio Longitudinal Study (2012–2022). Authors' elaboration. Note: Surrounding areas are 95% C.I.

References

1. Livingston, G., et al.: Dementia prevention, inter-vention, and care: 2020 report of the Lancet Commission. Lancet **396**, 413–446 (2020). https://doi.org/10.1016/S0140-6736(20)30367-6
2. Prince, M., Ali, G.-C., Guerchet, M., Prina, A.M., Albanese, E., Wu, Y.-T.: Recent global trends in the prevalence and incidence of dementia, and survival with dementia. Alzheimer Res. Therapy. **8**, 23 (2016). https://doi.org/10.1186/s13195-016-0188-8
3. Stern, Y.: What is cognitive reserve? Theory and research application of the reserve concept. J. Int. Neuropsychol. Soc. **8**, 448–460 (2002). https://doi.org/10.1017/S1355617702813248
4. Meng, X., D'Arcy, C.: Education and dementia in the context of the cognitive reserve hypothesis: a systematic review with meta-analyses and qualitative analyses. PLoS ONE **7**, e38268 (2012). https://doi.org/10.1371/journal.pone.0038268
5. Ricardo-Rodrigues, I., Jiménez-García, R., Hernández-Barrera, V., Carrasco-Garrido, P., Jiménez-Trujillo, I., López de Andrés, A.: Social disparities in access to breast and cervical cancer screening by women living in Spain. Public Health. **129**, 881–888 (2015). https://doi.org/10.1016/j.puhe.2015.02.021
6. Terraneo, M.: Inequities in health care utilization by people aged 50+: evidence from 12 European countries. Soc Sci Med **126**, 154–163 (2015). https://doi.org/10.1016/j.socscimed.2014.12.028
7. Bodryzlova, Y., Kim, A., Michaud, X., André, C., Bélanger, E., Moullec, G.: So-cial class and the risk of dementia: a systematic review and meta-analysis of the prospective longitudinal studies. Scand. J. Public Health. **51**, 1122–1135(2022). https://doi.org/10.1177/14034948221110019
8. Bougea, A., Anagnostouli, M., Angelopoulou, E., Spanou, I., Chrousos, G.: Psy-chosocial and trauma-related stress and risk of dementia: a meta-analytic systematic review of longitudinal studies. J. Geriatr. Psychiatry Neurol. **35**, 24–37 (2022). https://doi.org/10.1177/0891988720973759
9. Xu, W., et al.: Education and risk of dementia: dose-response meta-analysis of prospective cohort studies. Mol. Neurobiol. **53**, 3113–3123 (2016). https://doi.org/10.1007/s12035-015-9211-5

10. LaMantia, M.A., Stump, T.E., Messina, F.C., Miller, D.K., Callahan, C.M.: Emer-gency department use among older adults with dementia. Alzheimer Dis. Assoc. Disord. **30**, 35–40 (2016). https://doi.org/10.1097/WAD.0000000000000118
11. Motzek, T., Werblow, A., Tesch, F., Marquardt, G., Schmitt, J.: Determinants of hospitalization and length of stay among people with dementia – an analysis of statutory health insurance claims data. Arch. Gerontol. Geriatr. **76**, 227–233 (2018). https://doi.org/10.1016/j.archger.2018.02.015
12. Stirbu, I., et al.: Educational inequalities in avoidable mortality in Europe. J. Epidemiol. Commun. Health **64**, 913–920 (2010). https://doi.org/10.1136/jech.2008.081737
13. De Ronchi, D., et al.: Occurrence of cognitive impairment and dementia after the age of 60: a population-based study from Northern Italy. Dement. Geriatr. Cogn. Disord. **19**, 97–105 (2005). https://doi.org/10.1159/000082660
14. Di Carlo, A., et al.: Incidence of dementia, Alzheimer's disease, and vascular de-mentia in Italy. The ILSA Study. J. American Geriatr. Soc. **50**, 41–48 (2002). https://doi.org/10.1046/j.1532-5415.2002.50006.x
15. Bacigalupo, I., et al.: Identification of de-mentia and MCI cases in health information systems: An Italian validation study. Alzheimer's and Dementia Trans. Res. Clin. Interv. **8**, e12327 (2022). https://doi.org/10.1002/trc2.12327

Exploring Data Sharing Behaviour Using both Theory and Data-Driven Methods

Federica Cugnata[1,2(✉)], Chiara Brombin[1,2], Roberto Buccione[3], and Clelia Di Serio[1,2]

[1] University Center for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute S. Raffaele University, Via Olgettina 58, 20132 Milano, Italy
{cugnata.federica,brombin.chiara,diserio.clelia}@unisr.it
[2] Faculty of Psychology, Vita-Salute S. Raffaele University, Via Olgettina 58, 20132 Milano, Italy
[3] Vita-Salute S. Raffaele University, Via Olgettina 58, 20132 Milano, Italy
buccione.roberto@hsr.it

Abstract. Data sharing behavior is a complex phenomenon influenced by both individual and institutional factors, as well as the availability of internal resources and facilitating conditions. We study factors affecting data sharing by integrating two approaches: at first we test theory-driven relationships within a Structural Equation Modeling framework, then we explore data dependency structure within a data-driven perspective, using Bayesian Network models. The proposed approach is flexible and may be used to implement future intervention strategies aimed at enhancing data sharing practice.

Keywords: Data sharing · Bayesian Network · Structural Equation Modeling · Neo-institutional theory · Theory of Planned Behavior

1 Introduction

Data sharing represents a crucial aspect in ensuring the reproducibility of scientific findings. Although acknowledged by researchers as an essential practice for advancing scientific knowledge and progress, attitude towards data sharing may be affected by both individual and institutional factors as well as the availability of internal resources and facilitating conditions. Hence being a multifaceted construct, studying the complex interrelationships among variables involved in data sharing behaviour may be helpful to identify factors on which leverage to improve attitude towards this practice in research contexts. Capitalizing on theoretical models from the psychological literature, the Theory of Planned Behavior (TPB, [1]) and Neo-institutional theory [2] and adapting them to the specific context of data sharing attitude, we evaluate at first theory-driven relationships among relevant constructs. Actually, TPB highlights the role of attitude, subjective norms and perceived behavioral control in affecting behavioral intention and subsequent behavior. Neo-institutional theory emphasizes the role played

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 491–496, 2025.
https://doi.org/10.1007/978-3-031-64431-3_81

by the institutional environment, including rules and measures introduced by funding agencies and journals, norms and values established by the scientific community and the organizational culture [2] on individual actions. We propose and test a comprehensive model incorporating additional key factors expected to influence scientists' data sharing behavior, i.e., altruism and perceived behavioral control. After deriving PLS-SEM construct scores, we apply a data-driven approach to to deepen the knowledge about the interrelationships among psychological/behavioral constructs and individual factors, such as age, gender and tenure status.

2 Statistical Methods

The statistical analysis aimed at exploring the relationship between various factors influencing data sharing behavior. Our approach involves a staged methodology utilizing Structural Equation Modeling with Partial Least Squares estimation (PLS-SEM) and Bayesian Networks (BNs), each offering complementary features. The analysis progresses through two sequential steps. Initially, we apply PLS-SEM to test the hypothesized relationships among the collected variables based on the existing literature. PLS-SEM is a composite-based method to estimate structural equation models which goal is to maximize the explained variance of the latent constructs that are being considered in the model. This approach involves the development of two elements: a measurement model and a structural model ([5]). The measurement model explores into the relationship between questionnaire items and their respective latent constructs, while the structural model explores the connections between these latent constructs, allowing us to test hypotheses regarding direct and indirect effects. To assess the reliability and validity of the measurement instruments used to operationalize latent constructs, we employ a battery of measures including Cronbach's alpha, Composite Reliability (CR), Average Variance Extracted (AVE), and cross-loadings ([4–7]). PLS-SEM offers robust solutions for small sample sizes and complex models, accommodating missing values and various types of variables without distributional assumptions. Bootstrap methods are employed for statistical inference and the construct scores are derived by merging the scores of questionnaire items through a linear combination using the measurement model weights. Subsequently, these construct scores derived from PLS-SEM, are the basis for a Bayesian Network analysis [3,8] aiming at evaluating the relationships among individual factors. Bayesian Networks implement a graphical model structure called a directed acyclic graph, which comprises nodes representing random variables and arcs representing direct dependencies among these variables. They enable an effective representation and computation of a joint probability distribution over a set of random variables [3,8]. The purpose of utilizing Bayesian Networks in this research is to extract the dependence structure directly from the data, while excluding inconsistent variable directions. The network is estimated from data using a Tabu search greedy search algorithm, employing Akaike Information Criterion score functions. PLS-SEM models have been estimated using

SmartPLS [9] and the R package bnlearn [10] has been used to implement BN analysis.

2.1 Survey and Participants Description

Researchers from Vita-Salute San Raffaele University (UniSR) and I.R.C.C.S. San Raffaele Hospital were involved in the study. Data were collected through an online survey, developed in English, implemented and administered using the web-based platform Qualtrics. Demographics characteristics included age, gender, nationality, education, tenure status, primary discipline work and number of publications involving research data. To investigate data-sharing behavior, the same set of questions developed by Kim (2016, [11]) was administered, along with other constructs considered, including perceived behavioral control, altruism, and intention.

Complete questionnaires from 198 participants have been considered for this study. Most participants were females (n=121; 61%), with an average age of 42.76 years, Italian (n=185; 93.4%), tenured (n=97; 49.0%) and working in biomedical sciences (n=127;64%) and Medicine (n=73; 37%); mostly only in basic research (n=70; 44.6%) or in both clinical and basic (n=66; 42.0%), as expected considering the context of the research. Considering the number of publications involving actual research data per year, most of the subjects reported to publish between 1 and 2 papers per year (n=87; 43.9%) and 50 participants (25.3%) stated they publish 3–5 papers in the same period of time.

The data from the study indicates that data sharing behavior among surveyed researchers is not very frequent, with more than 40% of respondents reporting never depositing their data in a repository or public web space, or being personally asked to share data. Despite this, most respondents report a positive attitude toward data sharing, displaying willingness to share data with other researchers enhancing open science research. Although reporting to perceive more possible career benefits in data sharing practices than risks, they believe that data sharing requires time and effort. When contemplating regulatory pressures from funding agencies, journal publishers, and normative influences, the majority of participants perceive pressure to engage in data sharing behavior and view it as a standard expectation within the scientific community.

2.2 Structural Equation Modeling with Partial Least Squares Estimation

To evaluate the measurement model, we assess internal consistency, convergent validity, and discriminant validity of the research constructs. The values of Cronbach's alpha ranged from 0.74 to 0.94, exceeding the acceptable threshold of 0.7, indicating good internal consistency. Additionally, the Composite Reliability (CR) values range from 0.78 to 0.95, also surpassing the acceptable threshold of 0.7. Regarding convergent validity, the Average Variance Extracted (AVE) values range from 0.51 to 0.87, indicating satisfactory levels of convergent validity.

Furthermore, discriminant validity is confirmed through cross-loadings evaluation, where all indicators loaded higher on their respective constructs compared to other constructs.

The structural model associated with the hypothesized model is illustrated in Fig. 1, showing the path coefficients and corresponding p-values. Focusing on

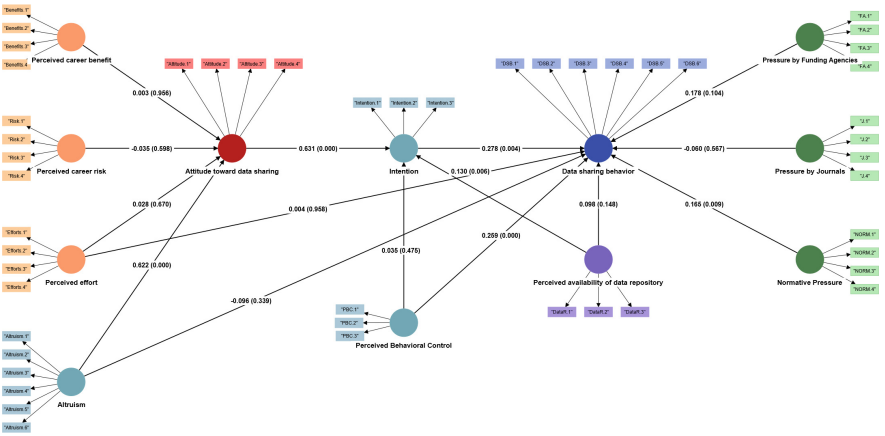


Fig. 1. PLS-SEM - Model. Path coefficients(p-values).

final outcome, PLS-SEM analysis highlights that intention positively and significantly influences data sharing behavior while attitude toward data sharing positively indirectly influenced data sharing behavior through Intention. Altruism positively and indirectly influences data sharing behavior through attitude. Moreover, data sharing behavior is positively and significantly influenced by perceived behavioral control and by normative pressure.

2.3 Bayesian Network Analysis

After the estimation of construct scores using PLS-SEM, Bayesian Networks are utilized to analyze the relationships between the considered constructs and individual characteristics. Initially, we eliminate blacklist arcs representing inconsistent variable directions. Subsequently, we construct network using the Tabu search greedy search algorithm. The resulting BN is depicted in Fig. 2. Demographic characteristics, education level, research work, the number of publications and tenure status influence data-sharing behaviour indirectly. It is noteworthy that, among others, tenure status impact on data-sharing through perceived pressure by funding agencies. Confirming the results of the PLS-SEM, the analysis of arc strength indicates that the most relevant relationships for data sharing behavior involve perceived behavioral control, intention, and normative pressure.

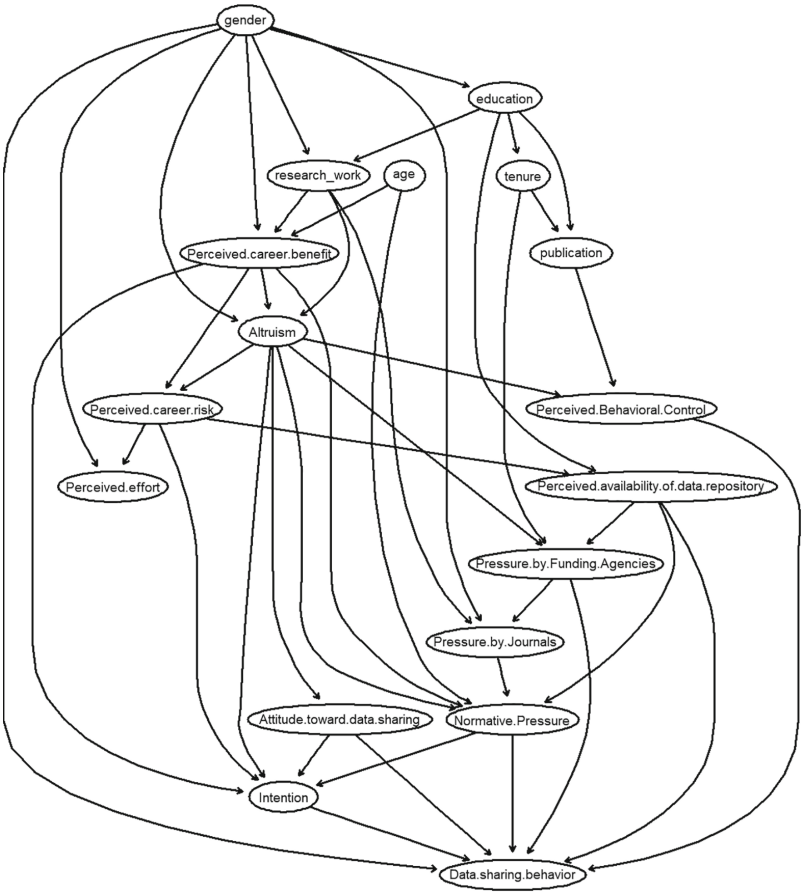


Fig. 2. BN structure

3 Conclusions

In this study we examined data sharing behaviour in a sample of researchers in the biomedical field. We explored the complex interrelationships among behavioural/psychological constructs and individual variables using both multidimensional theory-driven and data-driven frameworks. As matter of future research, leveraging on BN key features, alternative hypothetical scenarios will be analyzed, examining how changes in one or more variables propagates to other constructs and variables throughout the network. This scenario evaluation, conducted using BN analysis, will explore various configurations and assess their implications. Given the heterogeneity of data sharing behaviors and the highly flexibility of the proposed approaches, future studies should be implemented to

explore constructs of interest in other research areas. In this setting, BNs provide an appealing framework to compare dependency structures associated to different groups.

References

1. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**, 179–211 (1991)
2. Scott, W.R.: Institutions and organizations, 2nd edn. Sage Publications, Thousand Oaks, CA (2001)
3. Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic networks and expert systems: exact computational methods for bayesian networks. Springer Science & Business Media (2007). https://doi.org/10.1007/b97670
4. Hair, J.F., Sarstedt, M., Hopkins, L., Kuppelwieser, V.G.: Partial least squares structural equation modeling (PLS-SEM): an emerging tool in business research. Eur. Bus. Rev. **26**, 106–121 (2014)
5. Hair, J.F., Hult, T.M., Ringle, C.M., Sarstedt, M.: A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), SAGE Publications (2017)
6. Hair, J.F., Risher, J.J., Sarstedt, M., Ringle, C.M.: When to use and how to report the results of PLS-SEM. Eur. Bus. Rev. **31**, 2–24 (2019)
7. Hair, J.F., Hult, T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S.: Partial least squares structural equation modeling (PLS-SEM) using R–A workbook. Springer Nature (2022). https://doi.org/10.1007/978-3-030-80519-7
8. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, United Kingdom (2000)
9. Ringle, C.M., Wende, S., Becker, J.M. (2022). SmartPLS 4. Boenningstedt: SmartPLS. Retrieved from https://www.smartpls.com
10. Scutari, M.: Learning Bayesian networks with the bnlearn R package. J. Stat. Softw. **35**(3), 1–22 (2010). https://doi.org/10.18637/jss.v035.i03
11. Kim, Y.: Scientists' Data Sharing Behaviors. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor] (2016). https://doi.org/10.3886/E100087V7

Is Total Expenditure a Valid Proxy for Income in Statistical Matching? a Counterfactual Experiment

Elena Dalla Chiara[1(✉)], Marcello D'Orazio[2], and Federico Perali[3]

[1] Center of Economic Documentation (CIDE), University of Verona, Verona, Italy
elena.dallachiara@univr.it
[2] Department for Development of Methods and Technologies for Production and Dissemination of Statistical Information, Italian National Institute of Statistic (ISTAT), Roma, Italy
marcello.dorazio@istat.it
[3] Department of Economics, University of Verona, Verona, Italy
federico.perali@univr.it

Abstract. A critical but very common experimental situation in statistical matching (SM), where an expenditure survey is linked to a living conditions survey that includes an estimate of income, is the absence within the common variables as a proxy of income or of consumption. Researchers have often been in the situation of giving up the matching and looking for data that already contained the characteristics they were looking for. The aim of this study is to test whether total expenditure is a statistically reliable common variable for income when matching two datasets that do not include the income variable or a proxy of it. Our initial results derived in terms of matching quality produce evidence in support of the plausibility of total expenditure as a substitute for income.

Keywords: Income · Total expenditure · Statistical matching

1 Introduction

A critical but very common experimental situation in statistical matching (SM) where an expenditure survey is fused with a living conditions survey that includes an estimate of income, is the absence within the common variables of a proxy of income or of consumption. Researchers have often been in the situation of giving up the matching and looking for data that already containing the characteristics they were looking for, if any. The aim of this study is to test whether total expenditure is a statistically reliable variable for income when matching two data sets, such as the European Union Survey on Income and Living Conditions and the Household Budget Survey, where the latter one often does not include the income variable.

In the application of demand analysis, total expenditure is often seen as a legitimate proxy for income. If one is interested in measuring poverty, total household expenditure is often preferable. Slesnick (1993) points out that the choice between income and total

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 497–502, 2025.
https://doi.org/10.1007/978-3-031-64431-3_82

expenditure is crucial for poverty measurement. For the postwar period United States, Slesnick finds that poverty rates based on total expenditures are much lower than those based on income. The author attributes a possible explanation to consumption smoothing behavior. The permanent income hypothesis states that consumption decisions are based only on the permanent component of current income. For low-income household, this may be due to a large negative transitory income component. If we assume that consumption does not depend on transitory income losses, as suggested by the permanent income hypothesis, households with a temporarily low income should have a high consumption-to-income ratio. The evidence described by Slesnick underlines the important fact that the most disadvantaged households can be better identified if consumption rather than income is used as a measure of welfare.

Interestingly, the possibility of using total consumption as a proxy for income, and thus as an appropriate matching variable, is an option available to all matching practitioners, but one that has not yet been utilized because it has not been validated in the literature. We intend to fill this research gap and construct a counterfactual experiment using the 2019 wave of the US Consumer Expenditure (CE) survey of the Statistics Bureau, which includes total household expenditure and income. We simulate a real situation characterized by the absence of the treatment, where the treatment is the information on total household income, to compare these results with those where the treatment is present.

The goodness of the matching is evaluated in terms of targeting precision. We define measures based on the percentile distribution of both observed and imputed total expenditure. Statistical matching is performed using the hot deck (HD) and the propensity score (PS) methods to control for potential sources of bias when comparing the matching precision associated with the use of total expenditure or income. If the targeting precision obtained using either income or total expenditure applying both the HD and PS procedure is statistically comparable, then we can conclude that total expenditure is a good proxy for income and that consumption surveys can be reliably matched with income surveys when total expenditure is used as a common variable.

This note is structured as follows. First, we present the method used to set up the counterfactual experiment and define the measures of target accuracy. We then present the data and devote the following section to presenting the results regarding the quality of the matching procedure under the assumption that total expenditure is equal to actual income. The final section provides a preliminary answer to the question of whether total expenditure is a plausible proxy for income in statistical matching procedures that include consumption surveys without income information and income surveys without consumption information.

2 Method

This section describes the counterfactual matching experiment and the novel methodology to evaluate the matching quality.

Counterfactual Matching Experiment. We implement a "*homogeneous*" counterfactual experiment where the income variable of the donor data set (I_D) and the income variable of the recipient (I_R) are used as part of the matching variables and compare it with a "*hybrid*" experiment where I_D is not observed and substituted by total expenditure

in the donor (E_D), although we do not compare directly values of income with those of consumption but rather the relative position in the corresponding distribution. Holding all other matching variables and parameters fixed, the two experiments only differ in one matching variable, which we name "*hybrid*" variable. The hybrid variable takes the values of income in the recipient group and of expenditure in the donor group.

Both experiments are implemented by applying a nonparametric and a mixed SM approach to impute the household total expenditure in the recipient group. As a nonparametric approach, we implemented the random distance hotdeck (HD) method, and as a mixed approach, we used the propensity score (PS) method (D'Orazio et al. 2006, 2024; Dalla Chiara et al. 2023).

For the HD method, the set of matching variables includes the household (HH) size, the number of HH members with employment, and total income in the homogeneous experiment, and the hybrid variable for the experiment that uses total household expenditure as income. In this application, the units of recipient and donor are divided into groups obtained by crossing HH size with the number of employed HH members. Then, for each recipient unit in a given donation class, the total expenditure closest to the donor in terms of HH income or HH hybrid variable is imputed. Closeness is measured in terms of percentile distance estimated separately in each group.

The PS procedure is the probability of a unit being assigned to the treatment group based on their observed characteristics. The main objective of PS is to balance the distribution of the observed covariates between the recipient and donor groups, making them more comparable. The PS is estimated using a logistic regression model for the selected set of covariates common to both groups whose estimated score can be considered its synthetic indicator (Rässler 2002; Rosenbaum and Rubin 1983, 1985). In our application, a nearest neighbor (NN) algorithm with replacement, where one donor observation can be used more than once, is used to impute one unit in the recipient data set. The logit specification includes the following common variables: HH size, number of employed members in the HH, region of residence, education of the HH head, mortgage, age of the HH head, family type, and income or hybrid variable divided into twenty equal parts.

All SM procedures are performed using packages "*StatMatch*" (D'Orazio 2022) for HD and "*MatchIt*" (Ho et al. 2011) for PS, both available in the R statistical environment (R version 4.2.2; R Core Team 2023).

Matching Quality Evaluation. We evaluate the goodness of the matching procedure in terms of targeting precision. The imprecision is proportional to the probability of crossing the border of the percentile of the imputed case. For a sample of size n, each observation i belonging to a percentile $p = 1, .., P$ of the imputed empirical distribution $p_m(imputed)$ with the matching procedure has a correspondent position in the observed $p_o(observed)$ distribution. If the position in a given percentile of the imputed distribution is the same in the observed distribution, then the targeting error is zero. The difference between imputed and observed values measures the distance from an exact assignment. We define crossing as the event where the imputed position p_m differs from the observed position p_o.

Definition 1. **Crossing** (C) is the occurrence of a mismatching $p_m \neq p_o$.

The crossing either represents a move from the original (observed) position towards the bottom, reaching a pejorative living condition, or the top of the distribution where the living condition improves.

Definition 2. **Crossing Direction** (CD) is an indicator variable that informs about the direction of the crossing taking value of $\{-1, 0, 1\}$ according to the following rule:

$$\begin{cases} CD^{-} = -1 \; if \; p_m < p_o \\ CD = 0 \; if \; p_m = p_o \\ CD^{+} = 1 \; if \; p_m > p_o \end{cases}$$

when the value of CD is negative, then an observation of an imputed living condition (e.g., either relatively poor or rich) deteriorates its observed living condition. If CD equals 0, then there is no crossing, and the "poor" stays equally "poor" and the "rich" equally "rich". The case of a positive CD indicates an improvement in living conditions. We can now derive a measure of distance between percentile that is important to judge the quality of statistical matching.

Definition 3. **Percentile Mismatching** (PM) is measured as the difference between the imputed percentile $p_m(imputed)$ and the real percentile $p_o(observed)$ for each $i \in n$

$$PM_i = p_m(imputed_i) - p_o(observed_i) \; \forall \, i = 1, \ldots, n$$

where $PM_i = CD_i = 0$ when $p_m = p_o \; \forall p_m = 1, \ldots, P$ refers to a correct match.

3 Data

We use the 2019 wave of the Consumer Expenditure Survey (CE) administered by the US Census Bureau. The survey is nationally representative. The primary sample units (PSUs) are small clusters of counties grouped into geographic entities called "core-based statistical areas" (CBSAs). The survey collects data on expenditure, income, and demographics. The CE consists of two separate surveys: the Interview and Diary surveys. This study uses the Interview Survey, which includes 9,453 households (HHs).

To build the counterfactual experiment, the CE survey is split into recipient and donor groups of size equal to 40% and 60% of the total sample, respectively. The recipient group is drawn with stratified sampling with fixed number of units per each stratum defined according to income classes (decile of income as estimated in ASEC which is the main reference for estimating income) to achieve a sample where the distribution of income mimics that provided by the ASEC survey. The ASEC survey (Annual Social and Economic Supplement) is nationally representative and shares the same design of the CE. The donor group is selected with a simple random sample design from the total sample. We set up a simulation study by drawing fifty different recipient and donor groups and then match them with the proposed procedures.

4 Results

In this section, we present the results of the homogeneous and hybrid experiments implemented by adopting both the HD and PS statistical matching procedures. We judge the quality of this counterfactual experiment using the percentage of correct targets ($CD = 0$) and pejorative $\left(CD^{-}\right)$ and meliorative $\left(CD^{+}\right)$ crossing events. We also compute the proportion of mismatches by percentile (PM) in absolute value less or equal to 1, and the mean of the absolute value of PM. The latter information measures the percentage of making a small error and the mean targeting error.

Table 1 shows that the HD procedure performs better than PS when total expenditure is used as a proxy of total income. When looking at the results of the homogeneous experiment, the HD and PS procedures give comparable outcomes. Focusing only on the HD applications, when total expenditure is used as a proxy of total income, the performance of the procedure improves considerably. The exact matching based on the decile distribution of the total expenditure goes from 22% when income is account in both groups to 29.7% when total expenditure is considered as income. The probability of detecting a maximum error of ten percent varies from 52.7% in the homogenous case to 66.9% in the hybrid experiment. The mean of the difference in terms of decile of total expenditure distribution is 1.8 and 1.28 respectively.

Table 1. Comparison of the homogenous and hybrid counterfactual experiment controlling for the matching method using target distance measures.

	HD procedure		PS procedure	
Measures	*Homogeneous experiment*	*Hybrid experiment*	*Homogeneous experiment*	*Hybrid experiment*
Crossing events				
% of $CD = 0$	22.0	29.7	24.3	23.8
% of CD^-	39.1	33.1	36.1	35.8
% of CD^+	38.9	37.2	39.6	40.4
% $\|PM\| \leq 1$	52.7	66.9	51.7	53.3
Mean of $\|PM\|$	1.80	1.28	1.98	1.97

5 Conclusions

Our initial results show that the targeting precision measuring the matching quality obtained using either income or total expenditure applying both the HD and PS procedure is statistically comparable. We can therefore conclude that total expenditure is a good proxy for income and consumption surveys can be reliably matched with income surveys when total expenditure is used as a common variable.

References

Dalla Chiara, E., D'Orazio, M., Perali, F.: Statistical matching: hotdeck or propensity score?. In: Chelli, F.M., Ciommi, M., Ingrassia, S., Mariani, F., Recchioni, M.C.: Book of the Short Papers SIS 2023, pp. 611–616. Pearson (2023)

D'Orazio, M.: StatMatch: Statistical Matching or Data Fusion. R package version 1.4.1 (2022). https://CRAN.R-project.org/package=StatMatch

D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching. Wiley, Theory and Practice. Chichester (2006)

D'Orazio, M., Di Zio, M., Scanu, M.: Ask the experts: what is the state of play on statistical matching with a focus on auxiliary information, complex survey designs and quality issues? The Survey Statistician **89**, 47–58 (2024)

Ho, D.E., Imai, K., King, G., Stuart, E., Whitworth, A.: Nonparametric preprocessing for parametric causal inference. J. Stat. Softw. **42**(8), 1–28 (2011)

Rässler, S.: Statistical matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Springer, New York (2002)

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for casual effects. Biometrika **70**, 41–55 (1983)

Rosenbaum, P.R., Rubin, D.B.: Construction a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**(1), 33–38 (1985)

Slesnick, D.: Gaining gound: poverty in the Postwar United States. J. Polit. Econ. **101**(1), 1–38 (1993)

Exact Sampling of Two Coagulated Partitions

Marco Dalla Pria[1(✉)], Matteo Ruggiero[1], and Dario Spanò[2]

[1] University of Torino and Collegio Carlo Alberto, Turin, Italy
{marco.dallapria,matteo.ruggiero}@unito.it
[2] University of Warwick, Coventry, UK
d.spano@warwick.ac.uk

Abstract. Motivated by nonparametric posterior inference in a Bayesian problem where two independent unlabelled samples are available, we study the problem of sampling exactly in the set of coagulated partitions. These are obtained by matching a random number of cluster pairs chosen across the samples, and listing the remaining clusters unmatched. We propose an exact method for sampling from the correct target distribution, and present empirical evidence in comparison with the exact distribution.

Keywords: Bayesian inference · posterior distribution · random probability measure · unlabelled data

1 Introduction

A *partition* of $n \in \mathbb{N}$ is a vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{l(\gamma)}) \in \mathbb{N}^{l(\gamma)}$ such that $\gamma_i \geq \gamma_{i+1}$ and $|\gamma| = \sum_{i=1}^{l(\gamma)} \gamma_i = n$. We will call γ_i a *cluster* of γ. Consider now two such partitions $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{l(\gamma)})$ and $\omega = (\omega_1, \omega_2, \ldots, \omega_{l(\omega)})$, where $|\gamma|, |\omega|$ are not necessarily equal and the same holds for $l(\gamma), l(\omega)$. Now let the clusters in γ and ω be the nodes of a complete undirected bipartite graph $G = (V, E)$ with vertices $V = \{\gamma_1, \ldots, \gamma_{l(\gamma)}\} \cup \{\omega_1, \ldots, \omega_{l(\omega)}\}$ and edges E such that pairs (γ_i, ω_j) belong to E, but (γ_i, γ_j) and (ω_i, ω_j) do not.

Define a *matching* of γ and ω as a subset $M \subset E$ with the property that each node in V can be the endpoint of at most one edge in M. A matching induces a partition of the integer $|\gamma| + |\omega|$, henceforth referred to as *coagulation*, by coagulating the endpoints of each pair $(\gamma_i, \omega_j) \in M$ into a single cluster $\gamma_i \cup \omega_j$. A generic coagulation of γ and ω will be denoted $\mu = (\mu_1, \ldots, \mu_{l(\mu)})$, where by construction $\max\{l(\gamma), l(\omega)\} \leq l(\mu) \leq l(\gamma) + l(\omega)$, and each cluster μ_h can be given by either of these possibilities: γ_i for some i, ω_j for some j, $\gamma_i \cup \omega_j$ for some i, j. Hence all clusters in the original partitions that have not been matched are included in the coagulated partition μ in their original form. The set of all possible coagulations of γ and ω will be denoted $\mathrm{Coag}(\gamma, \omega)$.

In this contribution we are interested in the problem of sampling efficiently from the set $\mathrm{Coag}(\gamma, \omega)$. This problem is motivated by an inferential setting

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 503–506, 2025.
https://doi.org/10.1007/978-3-031-64431-3_83

in Bayesian nonparametrics whereby one has access to two different samples of size $|\gamma|$ and $|\omega|$, respectively, that have been collected independently but have been unlabelled, that is the only available information about the two samples is summarized in the observed groups sizes, represented by γ and ω. Posterior inference on the random probability measure which generated these partitions must then be necessarily based on the possible coagulations of γ, ω. Indeed, it is not known whether observations in γ_1 were of the same type as those in ω_j for a certain j, or in none of those. This inferential problem will be extensively treated in the manuscript in preparation [1].

The natural way to address this problem would be an exact computation of the probability of observing a specific coagulation if the matching are assigned uniformly at random. To understand this problem more specifically, denote by $\kappa(\mu|\gamma, \omega)$ the number of matchings of γ and ω that result in a given μ. Note that several matchings can produce the same coagulation. For example, define $\gamma = (3, 2, 2, 1)$ and $\omega = (2, 2, 1)$. A possible coagulation is $\mu = (4, 3, 2, 2, 2)$, which results from $\kappa(\mu|\gamma, \omega) = 6$ different matchings, since a cluster of size 4 can be produced by matching $(3, 1)$, or $(2, 2)$, and so on. On the other hand, there is a single matching yielding the coagulation $\mu = (3, 2, 2, 2, 2, 1, 1)$. In this example, there are $\sum_{\mu \in \text{Coag}(\gamma,\omega)} \kappa(\mu|\gamma, \omega) = 73$ possible matchings that yield $|\text{Coag}(\gamma, \omega)| = 20$ different coagulations.

Counting and sampling matchings exactly is a difficult problem arising in many other areas, like physics and chemistry in the context of the so-called *monomer-dimer systems*; cf. [3]. The cardinality κ of the set of matchings that produce each coagulation could in principle be computed by brute force through computer-based enumeration, but the problem becomes rapidly intractable as the cardinality of the sets of interest grows exponentially. Here we propose an alternative approach through a simple Monte Carlo algorithm that allows to sample coagulations from the correct distribution.

2 Exact Sampling of Two Coagulated Partitions

If we have a way to sample matchings uniformly at random, then by definition of κ each coagulation induced by these matchings is returned with probability proportional to $\kappa(\mu|\gamma, \omega)$. The empirical frequencies of the sampled coagulations would therefore be a Monte Carlo estimate of

$$g(\mu) = \frac{\kappa(\mu|\gamma, \omega)}{\sum_{\lambda \in \text{Coag}(\gamma,\omega)} \kappa(\lambda|\gamma, \omega)}, \tag{1}$$

which is the target distribution. The uniform distribution over matchings $M \subset E$ is therefore given by

$$\mathbb{P}(M) = \frac{1}{\sum_{\lambda \in \text{Coag}(\gamma,\omega)} \kappa(\lambda|\gamma, \omega)}. \tag{2}$$

Here we propose to sample conditionally on the length $l(\mu)$ of the induced coagulation, or, equivalently, conditionally on the number of matched pairs from the

original partitions, which can be written $d := l(\gamma)+l(\omega)-l(\mu)$. It turns out that there are

$$\sum_{\lambda\in\mathrm{Coag}(\gamma,\omega):\ l(\lambda)=l(\gamma)+l(\omega)-d} \kappa(\lambda|\gamma,\omega) = \binom{l(\gamma)}{d}\binom{l(\omega)}{d} d!$$

matchings coagulating exactly d pairs of clusters, and

$$\sum_{\lambda\in\mathrm{Coag}(\gamma,\omega)} \kappa(\lambda|\gamma,\omega) = \sum_{r=0}^{l(\gamma)\wedge l(\omega)} \binom{l(\gamma)}{r}\binom{l(\omega)}{r} r! \tag{3}$$

is the total number of matchings. These quantities induces the probability distribution on the number of matched pairs

$$\mathbb{P}(d) = \frac{\binom{l(\gamma)}{d}\binom{l(\omega)}{d} d!}{\sum_{r=0}^{l(\gamma)\wedge l(\omega)} \binom{l(\gamma)}{r}\binom{l(\omega)}{r} r!}, \quad d = 0, 1, \ldots, l(\gamma)\wedge l(\omega).$$

Now Eq. (2), by virtue of (3), becomes

$$\mathbb{P}(M) = \frac{1}{\sum_{r=0}^{l(\gamma)\wedge l(\omega)} \binom{l(\gamma)}{r}\binom{l(\omega)}{r} r!} = \frac{\binom{l(\gamma)}{d}\binom{l(\omega)}{d} d!}{\sum_{r=0}^{l(\gamma)\wedge l(\omega)} \binom{l(\gamma)}{r}\binom{l(\omega)}{r} r!}\mathbb{P}(M|d)$$

showing that the correct conditional distribution is

$$\mathbb{P}(M|d) = \frac{1}{\binom{l(\gamma)}{d}\binom{l(\omega)}{d} d!}, \qquad \text{for } |M| = d. \tag{4}$$

Equation (4) now suggests a straightforward strategy: first sample without replacement d clusters in each of γ and ω, then match randomly the chosen clusters across γ and ω.

Algorithm 1: Monte Carlo estimates of (1)

Input : two partitions γ, ω; number of draws N
$W(d) \leftarrow \binom{l(\gamma)}{d}\binom{l(\omega)}{d} d!$ for $d = 0, 1, \ldots, l(\gamma)\wedge l(\omega)$
for $i = 1, \ldots, N$ **do**
  sample d with probability $\propto W(d)$
  sample $M|d$ as described in (4)
  $\mu^{(i)} \leftarrow$ the coagulation induced by M
end
$f(\mu) \leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\mu^{(i)} = \mu)$ for all $\mu \in \cup_{i=1}^{N}\{\mu^{(i)}\}$
return $f(\mu)$ for all $\mu \in \cup_{i=1}^{N}\{\mu^{(i)}\}$

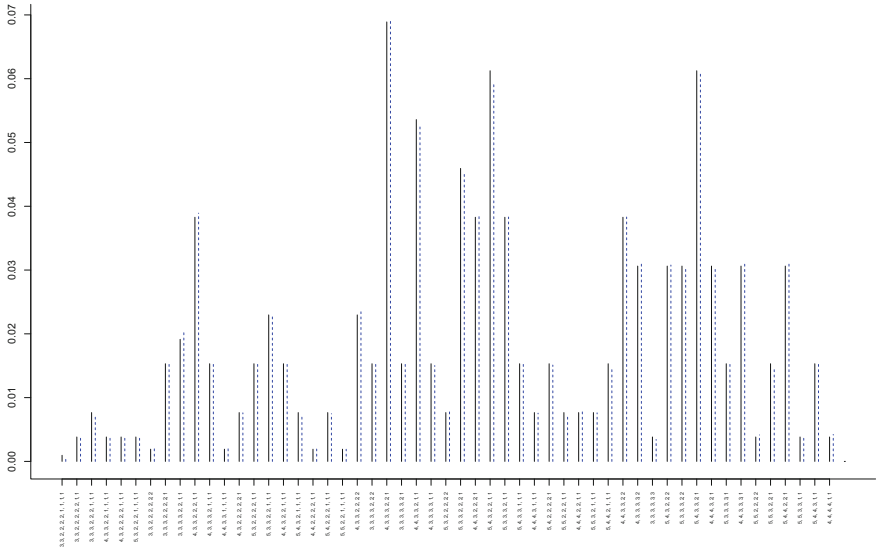


Fig. 1. Exact distribution (solid black) vs. Monte Carlo estimates (dashed blue) of (1) computed via Algorithm 1, based on $\gamma = (3, 3, 2, 2, 1, 1)$, $\omega = (2, 2, 1, 1)$, and $N = 10^5$ iterations. The points on the horizontal axis are the elements of $\text{Coag}(\gamma, \omega)$.

In summary, once d and $M|d$ are sampled, the coagulation μ induced by M is returned with probability proportional to $\kappa(\mu|\gamma, \omega)$ as in (1). We outline the pseudocode for this routine in Algorithm 1, and illustrate in Fig. 1 the empirical distribution of a Monte Carlo sample compared to the exact distribution for a small-dimensional problem with $\gamma = (3, 3, 2, 2, 1, 1)$ and $\omega = (2, 2, 1, 1)$.

References

1. Dalla Pria, M., Ruggiero, M., Spanò, D.: Inference for two-parameter Poisson–Dirichlet distributions based on multiple samples. In preparation (2024)
2. Griffiths, R.C., Ruggiero, M., Spano, D., Zhou, Y.: Dual process in the two-parameter Poisson-Dirichlet diffusion. arXiv (2024). URL http://arxiv.org/abs/2102.08520. ArXiv:2102.08520 [math]
3. Jerrum, M.: Mathematical foundations of the markov chain monte carlo method. In: Probabilistic methods for algorithmic discrete mathematics, pp. 116–165. Springer (1998). https://doi.org/10.1007/978-3-662-12788-9_4

A New Competition Model Combining the Lotka-Volterra Model and the Bass Model in Pharmacological Market Competition

Alessandra Dalla Valle(✉) and Cinzia Mortarino

Department of Statistical Sciences, University of Padova, Padua, Italy
alessandra.dallavalle@unipd.it

Abstract. The diffusion of products that compete in the marketplace is a strategic issue for market analysts. In this paper, we introduce two models that are essentially extensions of the class of the Lotka-Volterra competition models that have a consolidated history in the literature. The need to recover an excellent competition framework that was applied in innumerable disciplines, from biology to physics and so on, has essentially motivated this extension. The principal and appealing feature in these models is that they offer a definition of residual market that is more general with respect to the existing competition models available in literature, allowing that competing products contribute with different weights to perceived residual markets.

Keywords: Lotka-Volterra competition model · diachronic competition model · Diffusion processes

1 Introduction

The life cycle of a product in a target market is a strategic and fundamental marketing concept by firms that are planning new and innovative products take the place of older and outmoded ones. In concept, a product typically is launched, is adopted by consumers followed by a rapid growth period, reaches a deadlock, and then irremediably declines. These product phases may be quite different in both duration and framework. The univariate analysis of the life cycle of a single product had a consolidated theoretical development starting from the works of Fourt and Woodlock (1960), and Bass (1969). In particular, in the standard Bass model, the innovators, the first pioneering adopters, and the imitators (those who later adopted through word of mouth) are the principal latent drivers leading the diffusion process of a product in the market.

In the beginning of competition modeling, competition was considered a synchronic process, which is a very simplistic assumption if we consider that this condition is anything but common in real life. In fact, products typically enter the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 507–512, 2025.
https://doi.org/10.1007/978-3-031-64431-3_84

market at different times, and the roles that they assume reciprocally are quite different depending on whether they stood alone for a little while, stood alone for a long time, or stood in competition all along or only later in their life cycle. It is natural that the dynamics change, and a competition model should have a suitable degree of flexibility to allow for changing parameters. The diachronic competition approach is divided into balanced and unbalanced models. In the balanced models, word of mouth affects the adoptions of both products without distinctions, while in the unbalanced models, the interpersonal communication is split into two components: within-brand effects and cross-brand effects. The distinction appears crucial as it provides further flexibility because the products are perceived as different by consumers. In addition to the specific literature about competition models cited above, there is also a consolidated research line about competitive models that numbers the Lotka-Volterra model as one of the most widely used synchronic competition models. It was extensively used to model the dynamics of growth, coexistence, and survival of two (or even more) populations that share the resources in the same environment. It is essentially grounded on an alteration of a logistic equation to incorporate the interspecific competition coefficients, which, when added to the intraspecific coefficients make the modeling more complete. In this paper, we propose two extensions of the Lotka-Volterra competition model. The first one is inspired by the innovation component that is a fundamental part of the standard Bass model and permits one of the principal drivers of adoptions in the markets to be taken into account. The second extension allows that the Lotka-Volterra model may become diachronic by joining a standard Bass model that is able to capture the diffusion of the first competitor in its stand-alone period.

2 The Proposed Models

2.1 Lotka-Volterra Competition Models: A Brief Introduction

Lotka (1925) and Volterra (1926) introduced, independently, a mathematical competition model (LVC) that describes the dynamics of two species coexisting in the same environment and that act as prey and predator, competing for resources. The concept of environment and the dynamics of growth and death are intentionally simplified and, together with the assumption of no external special interventions result in a framework that it is still a convincing model. The differential equations of the LVC model are:

$$\begin{aligned} \frac{dX_1}{dt} &= \frac{\alpha_1 X_1}{K_1}(K_1 - X_1 - \alpha_{12} X_2) \\ \frac{dX_2}{dt} &= \frac{\alpha_2 X_2}{K_2}(K_2 - X_2 - \alpha_{21} X_1), \end{aligned} \tag{1}$$

where X_1 and X_2 are populations of species 1 and 2, respectively; α_1 and α_2 are the growth rates of species 1 and 2, respectively; K_1 and K_2 are the carrying capacities of the populations; α_{12} is the competition coefficient that measures

the competitive effect that population 2 has on population 1 and vice versa for α_{21}. In other words, α_{12} measures the external competition with respect to the internal one, that is, how many individuals of species 1 are equivalent to individuals of species 2. The interspecific coefficients identify the different influences that the two species reciprocally exercise and that condition intensively their whole competitive evolution. The system of Eq. (1) has no analytics solution, but approximate solutions are easily numerically computed.

2.2 The LVBC Diachronic Model

Our main idea is to obtain a new competition model that overtakes some limitations of LVC models (1) by introducing a framework that, on the one hand, emphasizes the role of innovators in the start-up of competition processes and that, on the other hand, considers that products competing in the market often are not synchronous. These extensions can be introduced separately and give rise to the models that we present in the following.

Let us consider two different brands competing in a market and sharing the same group of adopters. We propose an extension of model (1), denoted as LVBC model,

$$\begin{aligned} \frac{dX_1}{dt} &= \left(p_1 + \frac{\alpha_1 X_1}{K_1}\right)(K_1 - X_1 - \alpha_{12} X_2) \\ \frac{dX_2}{dt} &= \left(p_2 + \frac{\alpha_2 X_2}{K_2}\right)(K_2 - X_2 - \alpha_{21} X_1), \end{aligned} \tag{2}$$

where p_1 and p_2 are parameters that identify the innovation effect of the two different brands on the market. As in LVC model (1), X_1 and X_2 are cumulative adoptions of products 1 and 2, α_1 and α_2 are respectively the growth rates of product 1 and product 2, K_1 and K_2 are the carrying capacities of products 1 and 2, and α_{12} is the competition coefficient that measures the competitive effect that product 2 has on product 1, and vice versa for α_{21}. In other words, in the case of products instead of populations, α_{12} is interpreted as the reduction of growth in the diffusion of product 1 due to the presence of product 2 and vice versa for α_{21}. The competition between the two products can ultimately lead also to the exclusion of one of the products or to the coexistence of both products.

The contribution given in model (2) by innovation parameters p_1 and p_2, which represent a fundamental part of standard Bass models, is relevant because they start the mechanism of the diffusion process by highlighting the role of pioneers functioning by a subgroup of individuals. With respect to other competition models, model (2) extends the local perception of the residual market of each brand $(K_i - X_i), i = 1, 2$ through a quantity $\alpha_{ij} X_j$, with $i, j = 1, 2$, and $i \neq j$ that measures the strength of inhibition of the antagonist in the competition. However, because the class of Lotka-Volterra models is conceived for simultaneous processes, the application of model (2) cannot be considered completely satisfactory when two products enter the market at different times. Unfortunately, the condition of not equal launch times for products coexisting

in the market is a recurring case in practice and not only in particular cases or for special products.In the presence of a non trivial delay, it is not advantageous to avoid the market evolution of the first product because it implies a distorted estimation of competition between competitors. A possible solution is to modify the LVBC model by introducing a standard Bass model to capture the growth evolution in the stand-alone period for the first competitor. Because it is the initial period of a product's evolution, the choice of a simple Bass model does not appear simple, but is appropriate. Therefore, in diachronic competition, model (2) is modified as

$$\begin{aligned} \frac{dX_1}{dt} &= \left[\left(p_{1A} + \frac{\alpha_{1A}X_1}{K_{1A}}\right)(K_{1A} - X_1)\right] I_{[t<t_0]} + \\ &\quad + \left[\left(p_1 + \frac{\alpha_1 X_1}{K_1}\right)(K_1 - X_1 - \alpha_{12}X_2)\right] I_{[t\geq t_0]} \\ \frac{dX_2}{dt} &= \left(p_2 + \frac{\alpha_2 X_2}{K_2}\right)(K_2 - X_2 - \alpha_{21}X_1), \end{aligned} \tag{3}$$

where p_{1A}, α_{1A}, and K_{1A} are respectively the coefficients of innovation, imitation, and potential market for the first product in the stand-alone period until time t_0 when a second competitor gets in competition with the first already existing product, and I_η indicates a dummy function that is equal to 1 when condition η is satisfied.

This model, denoted by LVBC-diac, recovers those data that are normally censored using a synchronic modeling and that, on the contrary, are essential to describe the whole evolution of the product first coming to the market. Furthermore, this clearly affects the final differential equations system by returning a more reliable competition model.

Because LVBC-diac (3) and also LVC (1) and LVBC (2) models do not have closed-form solutions, they have to be applied directly beginning with the differential form, which will be used to estimate the parameters involved.

3 A Real Data Application

The need for testing the performance of our model with respect to the preceding models introduced in the literature drove us to make comparisons using the same data as Guseo and Mortarino (2012) and Guseo and Mortarino (2014). Data were provided by IMS-Health, Italy, and consist of cumulative quarterly number of packages of Cimetidine and Ranitidine sold in Italy. Figure 1 shows the correspondence between the observed trends of the number of packages sold for each pharmaceutical drug and the values fitted by the LVBC-diac model whose parameter estimates are summarized in Table 1. The estimated coefficient of innovation p_1 is evidently and correctly greater than p_2 given that the innovation process was started just by Cimetidine when Ranitidine was missing. Moreover, the relevant weight in the competition is confirmed also by the different values of the competition coefficients α_{12} and α_{21}, proving the inhibition strength of

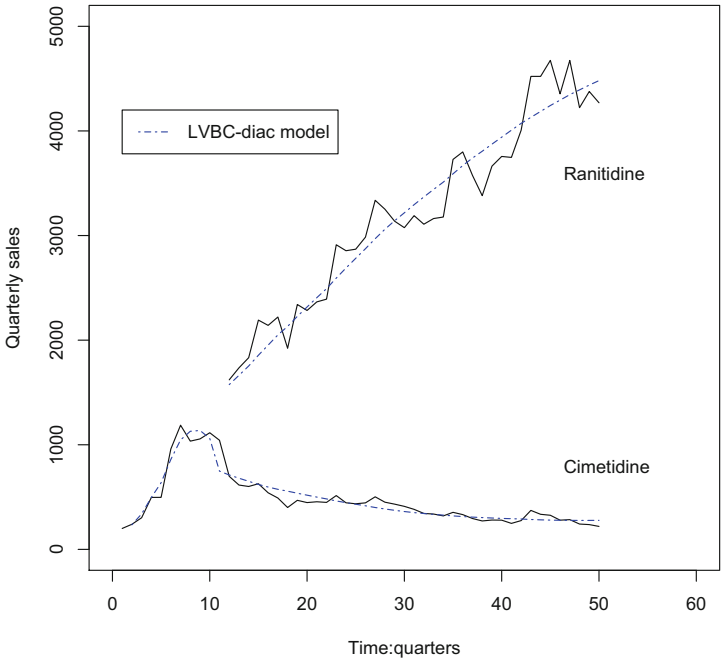


Fig. 1. The LVBC-diac model: observed versus fitted values for Cimetidine and Ranitidine

Table 1. Multivariate estimation results for model LVBC-diac.

Parameters	Estimate	Standard Error
p_{1A}	0.00640	0.01099
α_{1A}	0.39230	0.09459
K_{1A}	11280	2067
p_1	0.03246	6.894
α_1	-0.03374	7.483
K_1	38280	8157000
α_{12}	0.05714	1.027
α_2	0.09758	0.03641
p_2	0.000302	0.00192
K_2	4803000	30220000
α_{21}	90.47	616.7

Ranitidine. The fit of model (3) seems quite good, but it has to be assessed by also evaluating other existing models in literature.

Figure 2 shows that, for Cimetidine, (left side), the LVBC-diac model is perfectly intermediate between the other two models and seems to follow the

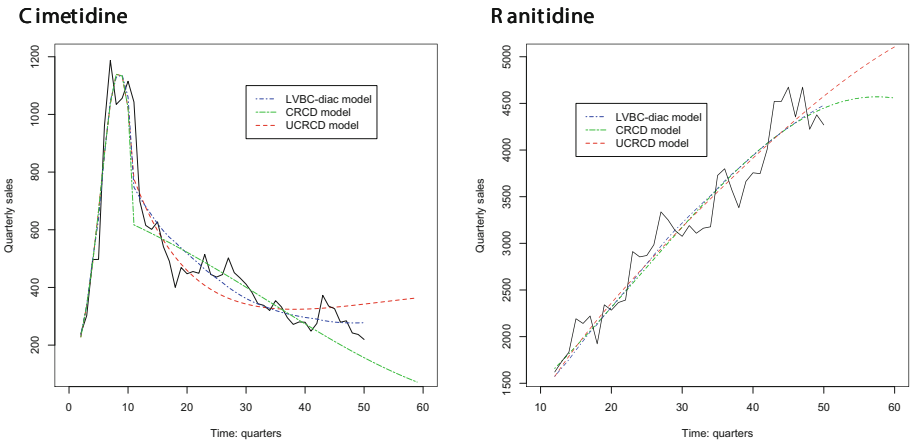


Fig. 2. Comparison between observed and fitted values for the LVBC-diac model, CRCD model, and UCRCD model with respect to Cimetidine (left side) and Ranitidine (right side)

observed trend better, especially in the final period where CRCD and UCRCD take different directions. In the initial part, the three models are quite perfectly overlapping given that they all lie on the same Bass model, but afterwards, the difference is quite evident. For Ranitidine, (right side), the differences are notably reduced, and the curves are overlapped, so the trend is notably good for each of the considered models.

References

Bass, F.M.: A new-product growth model for consumer durables. Manage. Sci. **15**, 215–227 (1969)

Fourt, L.A., Woodlock, J.W.: Early prediction of market success for grocery products. J. Mark. **25**, 31–38 (1960)

Guseo, R., Mortarino, C.: Sequential market entries and competition modelling in multi-innovation diffusions. Eur. J. Oper. Res. **216**(3), 658–667 (2012)

Guseo, R., Mortarino, C.: Within-brand and cross-brand word-of-mouth for sequential multi-innovation diffusions. IMA J. Manag. Math. **25**(3), 287–311 (2014)

Lotka, A.J.: Elements of Physical Biology, Williams and Wilkins (1925)

Volterra, V.: Variations and Fluctuations of The Number of Individuals in Animal Species Living Together. McGraw-Hill, Animal Ecology (1926)

Measuring Sustainability in Tourism via a Multi Criteria Decision Making Approach

G. Damiana Costanzo[1], Barbara Dattilo[2], Mascia Di Torrice[2], Michelangelo Misuraca[1(✉)], and Maria Spano[3]

[1] DiScAG - University of Calabria, Rende, Italy
{damiana.costanzo,michelangelo.misuraca}@unical.it
[2] Italian Institute of Statistics (Istat), Rome, Italy
{maditorr,dattilo}@istat.it
[3] DiSES - University of Naples Federico II, Rende, Italy
maria.spano@unina.it

Abstract. Tourism sustainability has to fully account for economic, socio-cultural and environmental aspects, addressing the needs of visitors and host communities. As this concept has a multidimensional nature, it is necessary to build a synthetic indicator that is useful for assessing Tourism sustainability at a territorial level. One of the critical problems is the choice of a suitable set of weights for the different considered sub-indicators. This work uses a weighting strategy based on Stochastic Multi-attribute Acceptability Analysis to deal with this task. We employed the synthetic indicator of Tourism sustainability obtained with this procedure to rank the 20 Italian regions, showing the preliminary results of an ongoing study on the phenomenon.

Keywords: synthetic indicator · SMAA · regional development

1 Introduction

In the last few years, scholars and policy-makers have intensively debated the importance of sustainability in Tourism. The basic idea is that the development of Tourism from an economic perspective has to be sensible to the interest of all the potential stakeholders, including tourists themselves as well as the host communities, ensuring the satisfaction of travellers and day-trippers but mainly respecting and preserving local cultures and natural and social diversity [5].

Several solutions for a quantitative assessment of Tourism sustainability at a territorial level have been discussed in the literature. There is substantial agreement on the multidimensionality of the concept [2] and its key dimensions, involving the economic, socio-cultural, and environmental spheres. However, there are different proposals on the methodological viewpoint. Leaving aside the problem of selecting a set of elementary indicators covering the different facets of the involved dimensions (which reflect the theoretical framework underlying the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 513–518, 2025.
https://doi.org/10.1007/978-3-031-64431-3_85

concept to measure) and a suitable aggregation function for building a synthetic indicator, one of the most tricky aspects is choosing a weighting system that numerically expresses every single indicator importance before aggregating them into an overall evaluation [7]. Typically, weights assigned to the indicators can be based on a public/expert opinion or statistically determined. However, in both cases, the adopted scheme significantly affects the aggregation. For example, if a synthetic indicator is used for territorial comparisons (e.g., to look at the competitiveness of those areas), adopting different weights could change the resulting ranking and induce a different interpretation of the phenomenon.

Since none of the proposed solutions is free from potential hazards [3], here we implement a solution for building a Tourism Sustainability Index based on a *multi-criteria decision making* (MCDM) approach [4,10]. Multi-criteria analyses solve complex problems by assessing all the indicators observed on a set of entities with missing weight information or, more generally, when the corresponding weights are subject to uncertainty. In particular, our strategy refers to the *Stochastic Multi-attribute Acceptability Analysis* (SMAA) [6,9], a method that considers the complete weighting space for the involved single indicators. Each vector of non-negative weight is randomly extracted under the hypothesis of uniform probability distribution. The "objectivity" of the adopted weighting scheme is given by considering a considerable sample of weights' vectors and deriving a ranking of the analysed entities based on a consensus among the overall indicators obtained by applying the different weights' vectors. An application on the sustainability assessment of Tourism in Italy is used to show the proposal's effectiveness, ranking Italian regions with respect to a range of economic, socio-cultural, and environmental indicators commonly employed in the literature to highlight the diverse facets of the phenomenon.

2 Materials and Methods

To evaluate the Tourism sustainability of the Italian regions, we collected 32 sub-indicators concerning economic (9 criteria, including the quality and quantity of employment in Tourism and the impact of Tourism on regional expenditure), socio-cultural (9 criteria, including the impact of tourists on residents and territories, and tourists' satisfaction about services), and environmental (14 criteria, including air quality, impact of tourists on waste management and freshwater uses) aspects. The data were collected from several sources, such as the *Italian Institute of Statistics* (Istat), the *Italian Institute for Environmental Protection and Research* (ISPRA), the *Bank of Italy*, the *Territorial Public Accounts* (CPT) system, the *Italian Federation of Hotel and Tourism Associations* (Federalberghi). All the sub-indicators reflect the different aspects at a regional level in the period 2020–2022. More details on the data will be shown elsewhere.

Before performing the analysis and constructing the synthetic indicator, each sub-indicator was normalised to ensure a common measurement scale and then polarised to ensure all criteria were going in the same direction of the phenomenon (i.e., higher values for higher sustainability). The weights of the criteria were determined through SMAA, using a 1 million vector uniform sampling.

2.1 Stochastic Multi-attribute Acceptability Analysis

In determining how the various aspects may characterise Tourism sustainability at a territorial level, it is possible to take into account the whole spectrum of preferences and attitudes towards the studied issue. Scholars, for example, may be interested in stressing the socio-cultural or environmental aspects, whereas policy-makers may focus more on economic performance. From a comparative perspective, it is then reasonable to expect that some areas would be more valued by some aspects while different ones would value other areas. This diversified appreciation of the sustainability facets leads to the choice of different weights assigned to the sub-indicators that are aggregated to obtain an overall measure. Any single weights vector allows building a synthetic indicator that, at best, represents a satisfactory outcome only for a fraction of individuals, affecting in this way the representativeness of the achieved assessment. From a methodological standpoint, this plurality of weights vectors relieves robustness problems [8].

The SMAA approach considers the whole set of possible weights to be assigned to a set of sub-indicators, approximated by an extensive sample of randomly extracted weight vectors. In MCDM problems, a set A of n alternatives (units) is evaluated on a set C of p criteria to deal with tasks such as choosing the best alternative or ranking all alternatives from the best to the worst. A weighted sum is the function most commonly employed to aggregate the evaluations of the different criteria. After assigning a weight $w_j \in \Re_+^p$ to each criterion $c_j \in C$ (with $j = 1, \ldots, p$), stated that $w_1 + \cdots + w_p = 1$, each alternative $a_i \in A$ (with $i = 1, \ldots, n$) has the following overall evaluation:

$$\mathrm{SI}_{a_i;\mathbf{w}} = \sum_{j=1}^{p} w_j c_j(a_i) \tag{1}$$

where w_j reflects the importance given to the *j-th* criterion, and $c_j(a_i)$ is the value observed for the criterion in the alternative a_i. As stated above, the key problem is that the criteria importance may vary among individuals, implying that not one single best weights vector $\mathbf{w}$ represents the entire population. Alternatives ranking changes as the importance of criteria changes.

To embody unknown preferences on the weights, SMAA considers a probability distribution in the set W of possible weights, with density function $f_W(w)$, and a probability distribution in the χ evaluation space, i.e. the space of the value that can be taken by the criteria c_j, with density function $f_\chi(\xi)$. The lack of knowledge concerning the weights can be expressed by a uniform weight distribution in W. To rank the alternatives (from 1 to n) according to Eq. 1, it is possible to define a ranking function for each a_i:

$$\mathrm{rank}(i, \xi, \mathbf{w}) = 1 + \sum_{i' \neq i} \rho[\mathrm{SI}_{a_{i'};\mathbf{w}} > \mathrm{SI}_{a_i;\mathbf{w}}] \tag{2}$$

where $\rho(false) = 0$ and $\rho(true) = 1$. In practice, the rank of a_i, given a weights vector $\mathbf{w}$, is one plus the number of alternatives performing worse than the a_i

on the basis of the selected criteria. The approach, in particular, relies on the analysis of the sets of favourable rank weights $W_i^r(\xi)$ defined as:

$$W_i^r(\xi) = \{\mathbf{w} \in W : \text{rank}(i, \xi, \mathbf{w}) = r\} \tag{3}$$

with $\mathbf{w} \in W_i^r(\xi)$ guaranteeing that the *i-th* alternative obtains a rank r. From Eq. 3, it is possible to compute the so-called *rank acceptability index* (RAI), that is the expected volume of the set of favourable rank weights:

$$b_i^r = \int_{\chi} f_{\chi}(\xi) \int_{W_i^r(\xi)} f_W(w)\ dwd\xi \tag{4}$$

The RAI can be interpreted as the probability that a_i has rank r, that is, the ratio of the number of weights vectors by which the *i-th* alternative gets the *r-th* position in the ranking to the total amount of possible weights vectors.

In this study, we applied the aforementioned approach to obtain a synthetic indicator of Tourism sustainability, ranking the 20 Italian regions (the alternatives) according to a set of economic, socio-cultural and environmental sub-indicators (the criteria). All the results presented here have been obtained using the SMAA method implemented in the Python library crispyn [1].

3 Preliminary Results and Future Research Directions

The ranking related to the selected sub-indicators contains a large amount of information on regional Tourism sustainability facets, which aligns with the idea of the multidimensionality of the phenomenon widely assumed in the literature.

Table 1 displays the *alternative* labels A_i assigned to the 20 Italian regions (the order follows the numeric codes officially assigned by Istat), making the results reported hereunder more easily readable.

Table 1. Labels of the 20 Italian regions

A_i	Region	A_i	Region	A_i	Region	A_i	Region
A_1	Piedmont	A_6	Friuli-Ven.Giulia	A_{11}	Marche	A_{16}	Apulia
A_2	Aosta Valley	A_7	Liguria	A_{12}	Lazio	A_{17}	Basilicata
A_3	Lombardy	A_8	Emilia-Romagna	A_{13}	Abruzzo	A_{18}	Calabria
A_4	Trentino-S.Tyrol	A_9	Tuscany	A_{14}	Molise	A_{19}	Sicily
A_5	Veneto	A_{10}	Umbria	A_{15}	Campania	A_{20}	Sardinia

To compute the RAIs for each region, the ratio between the occurrences a region achieves a given rank and the total number of cases are considered. Figure 1 depicts as a heat-map the matrix with the RAIs of the 20 regions. Each row of the matrix contains the indexes for each possible rank. It is also possible to obtain the corresponding *central weight vectors*, i.e. the barycentre of the set of weight vectors for which a region is the best alternative among the others.

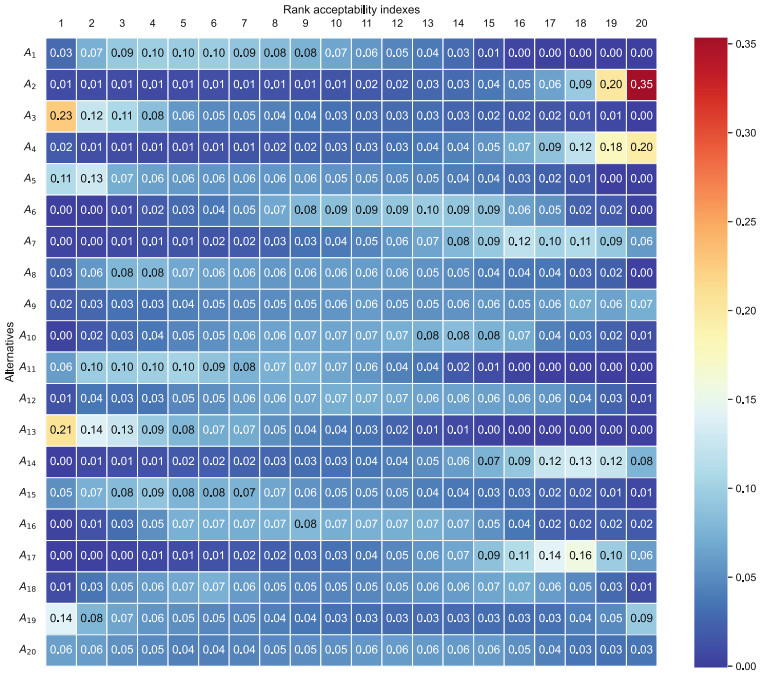


Fig. 1. Rank acceptability indexes for the 20 regions

It is possible to see that Aosta Valley (A_2) and Trentino-S.Tyrol (A_4) reached the highest RAI for the 19th and 20th positions in the ranking, revealing the worst level of Tourism sustainability. These regions suffered, in particular, the impact of Tourism on freshwater use and waste management, as well as a higher level of pollution caused by on-road tourists' travels. Noteworthy, both regions are located in the North of Italy and are mainly mountain territories. On the other side, Lombardy (A_3) and Abruzzo (A_{13}) reached the highest RAI for the 1st position in the ranking, revealing the best level of Tourism sustainability. Abruzzo, in particular, showed higher weights for tourist satisfaction concerning entertainment and food service, and for land conservation areas.

Table 2. Ranking of 20 regions provided by the SMAA ranking function

$\mathbf{A_i}$	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}
Rank	4	20	2	19	5	15	16	7	15	13	3	12	1	17	6	10	18	11	8	9

Table 2 displays the final ranking of the regional Tourism sustainability according to the 32 criteria, employing Eq. 2 with regional central weight vectors. It is possible to see that the smallest regions and most peripheral regions (Aosta Valley, Trentino-S.Tyrol, Molise, Basilicata) experienced the lowest level on the sustainability indicator, whereas bigger Northern regions (Lombardy, Piedmont

and Veneto) suffered less the impact of Tourism on residents and territories, experiencing the highest level on the sustainability indicator. This polarisation of regions located in different parts of Italy has already been observed in other studies, and it is interesting to explore by means of additional analyses.

Our findings are influenced by the conceptual framework underlying the indicator obtained through the SMAA approach, remembering that the COVID-19 pandemic strongly influenced Tourism in 2020–2021. Even if the sub-indicators employed here have already been included in other studies on Tourism sustainability, it is necessary to check the influence of single criteria on the overall measure and the impact on the resulting ranking. This deeper understanding of the determinants of Tourism sustainability will be paired with a comparative study of the other weighting schemes used to construct synthetic indicators, as well as performing comparative analyses with other regions or countries.

Moreover, trends in Tourism sustainability over time will be considered by conducting longitudinal analyses. Tracking changes in sustainability performance may help revealing emerging patterns and support public and private strategies.

References

1. Bączkiewicz, A., Wątróbski, J.: Crispyn - a Python library for determining criteria significance with objective weighting methods. SoftwareX **19**, 101166 (2022). https://doi.org/10.1016/j.softx.2022.101166
2. Fernández, J.I.P., Rivero, M.S.: Measuring tourism sustainability: proposal for a composite index. Tour. Econ. **15**, 277–296 (2009). https://doi.org/10.5367/000000009788254377
3. Greco, S., Ishizaka, A., Tasiou, M., Torrisi, G.: On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc. Indic. Res. **141**(1), 61–94 (2018). https://doi.org/10.1007/s11205-017-1832-9
4. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Wiley, New York (1976)
5. Ko, T.G.: Development of a tourism sustainability assessment procedure: a conceptual approach. Tour. Manag. **26**, 431–445 (2005). https://doi.org/10.1016/j.tourman.2003.12.003
6. Lahdelma, R., Hokkanen, J., Salminen, P.: SMAA - stochastic multiobjective acceptability analysis. Eur. J. Oper. Res. **106**, 137–143 (1998). https://doi.org/10.1016/S0377-2217(97)00163-X
7. Mikulić, J., Kožić, I., Krešić, D.: Weighting indicators of tourism sustainability: a critical note. Ecol. Indic. **48**, 312–314 (2015). https://doi.org/10.1016/j.ecolind.2014.08.026
8. Saisana, M., Saltelli, A., Tarantola, S.: Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. J. R. Stat. Soc. Ser. A Stat. Soc. **168**, 307–323 (2005). https://doi.org/10.1111/j.1467-985X.2005.00350.x
9. Tervonen, T., Lahdelma, R.: Implementing stochastic multi-criteria acceptability analysis. Eur. J. Oper. Res. **178**, 500–513 (2007). https://doi.org/10.1016/j.ejor.2005.12.037
10. Von Winterfeldt, D., Edwards, W.: Decision Analysis and Behavioral Research. Cambridge University Press, New York (1986)



Biological Species Delimitation Based on Genetic and Spatial Dissimilarity Using Partial Mantel Tests

Gabriele d'Angella(✉) and Christian Hennig

Alma Mater Studiorum - University of Bologna, Via Belle Arti, 41 -, 40126 Bologna, (BO), Italy

gabriele.dangella2@unibo.it, christian.hennig@unibo.it

Abstract. The delimitation of biological species, i.e., deciding which individuals belong to the same species and whether and how many different species are represented in a data set, is key to the conservation of biodiversity. Much existing work uses only genetic data for species delimitation, but this can be misleading, because geographically distant groups of individuals can be genetically quite different even if they belong to the same species. This paper investigates the problem of testing whether two potentially separated groups of individuals can belong to a single species or not based on genetic and spatial data. A way to do this is via partial Mantel tests (PMTs), in which significance is traditionally assessed through permutations. By replacing these with jackknife and bootstrap, two new versions of PMT for species delimitation are introduced in this study, where their type-I-error rate and power are compared with permutation-based PMT on data generated by `GSpace`, a spatially-explicit genetic data simulator. Results suggest that bootstrap-based PMTs have unacceptable type-I-error, while jackknife-based ones are more conservative than permutation-based ones, but less powerful.

Keywords: species delimitation · distance-distance methods · partial Mantel tests · `GSpace`

1 Introduction

Species delimitation is key for the preservation of biodiversity and consists in determining how many biological species are represented in a dataset. To this aim, non-spatial molecular methods are often used, but neglecting geographic information when delimiting species can lead to the overestimation of the genetic structure in a dataset [5]. This is more likely to happen when individuals in a species are isolated by distance (IBD), that is, their genetic differences increase with their geographic separation. A way to tackle this problem is to delimit species through methods that model the relationship between genetic and geographic distance. Being based on simulated data, this study uses the Euclidean distance, but more sophisticated geographic distances exist [8]. Consider a setup

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 519–524, 2025.
https://doi.org/10.1007/978-3-031-64431-3_86

with two putative groups of individuals to be tested for conspecificity: for every pair of individuals in the sample, a grouping dissimilarity is computed, that takes value 1 when they belong to different groups and 0 otherwise. The impact of grouping on the genetic dissimilarity can be quantified controlling for the effect of geographic distance: if the genetic structure is positively associated with the grouping dissimilarity even after accounting for the geographic distance, there is evidence that the two groups represent distinct species, since their molecular differences cannot be explained by geographic separation. This idea was applied by [9], who resorted to permutation-based partial Mantel tests (PMTs) to assess the significance of the partial correlation coefficient between genetic and grouping dissimilarities given the geographic distance. This partial correlation analysis is carried out through permutations because dissimilarity data is dependent by construction, making standard inference for partial correlations not applicable. Alternative approaches to the permutations include jackknife and bootstrap [2]: the latter is discouraged in distance-distance methods because usually introduces more null distances than those actually present in the data, but no systematic study to date investigated its performance for species delimitation; as for the jackknife, to the best of our knowledge, no study ever paired it with PMT for species delimitation. This study compares the type-I-error rate and power of these three versions of the PMT via spatially-explicit genetic simulations performed by `GSpace` [10].

2 Data

Spatially-explicit genetic data consists of individuals carrying information about their location and genetic make-up. Methods will be applied on n individuals from two groups, with known membership. Hence, two columns will correspond to the unit's coordinates, one column will report the group labels (either group 1 or 2) and the other P columns will be loci (locations on the DNA). Diploid loci contain two alleles, each of which is represented as a character: alleles like "B" [6] form diploid loci like "BA". The row vector representing the i^{th} individual is denoted $\mathbf{z}_i^\intercal = (z_i^x, z_i^y, z_i^c, Z_i^1, \cdots, Z_i^P)$, where each observed locus Z_i^p, $p = 1, \ldots, P$, is a set of characters. The Euclidean distance $d_x(\mathbf{z}_i, \mathbf{z}_j) = \sqrt{(z_i^x - z_j^x)^2 + (z_i^y - z_j^y)^2}$ is employed as geographic distance, whereas the shared allele dissimilarity [1] is used as genetic dissimilarity:

$$d_y(\mathbf{z}_i, \mathbf{z}_j) = 1 - \frac{1}{2P} \sum_{p=1}^{P} \left| Z_i^p \cap Z_j^p \right| \cdot \left[1 + \mathbb{1}\left(|Z_i^p| + |Z_j^p| = 2 \right) \right],$$

where $\mathbb{1}(\text{condition}) = 1$ if the condition is true and zero otherwise. The grouping distance is computed as $d_g(\mathbf{z}_i, \mathbf{z}_j) = \mathbb{1}\left(z_i^c \neq z_j^c\right)$.

Assume to have $n_k = |\{i : z_i^c = k\}|$ individuals belonging to cluster k, with $k = 1, 2$ and $n = n_1 + n_2$: these represent the two candidate (or putative) species to be tested for conspecificity. The resulting $n(n-1)/2$ geographic distances in

the dataset can be stored in the lower-triangular part, denoted $\{d_x(\mathbf{z}_r, \mathbf{z}_c)\}_{c<r\leq n}$, of a block matrix $\mathbf{D}_x = \left(\begin{array}{c|c} \mathbf{D}_x^{11} & \mathbf{D}_x^{12} \\ \hline \mathbf{D}_x^{21} & \mathbf{D}_x^{22} \end{array}\right)$, where matrix $\mathbf{D}_x^{11}$ stores the distances among the observations belonging to group 1, $\mathbf{D}_x^{22}$ those within group 2 and $\mathbf{D}_x^{12} = (\mathbf{D}_x^{21})^\intercal$ those among individuals of different groups. Likewise for the genetic and grouping distances in $\mathbf{D}_y$ and $\mathbf{D}_g$.

3 Partial Mantel Test

Let the linear correlation between the genetic and geographic distances in the dataset be:

$$r(\mathbf{D}_y, \mathbf{D}_x) = \frac{\sum_{c<r\leq n}(d_y(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_y)(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x)}{\sqrt{\sum_{c<r\leq n}(d_y(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_y)^2 \sum_{c<r\leq n}(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x)^2}},$$

where $\bar{d}_y = \frac{1}{w}\sum_{c<r\leq n} d_y(\mathbf{z}_r, \mathbf{z}_c)$ is the average genetic dissimilarity and $\bar{d}_x = \frac{1}{w}\sum_{c<r\leq n} d_x(\mathbf{z}_r, \mathbf{z}_c)$ is the average geographic distance, with $w = n(n-1)/2$. The test statistic of the PMT is:

$$r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x) = \frac{r(\mathbf{D}_y, \mathbf{D}_g) - r(\mathbf{D}_y, \mathbf{D}_x)r(\mathbf{D}_g, \mathbf{D}_x)}{\sqrt{(1 - r(\mathbf{D}_y, \mathbf{D}_x)^2)(1 - r(\mathbf{D}_g, \mathbf{D}_x)^2)}}, \tag{1}$$

and under the null hypothesis its true value in the population is lower than or equal to zero. Under the alternative hypothesis, any positive association between genetic and geographic distances (IBD behaviour) is not sufficient to explain the genetic discrepancies between individuals from two putative species and there is evidence that they may constitute distinct species. Hypothesis testing is carried out by permuting the rows and corresponding columns of $\mathbf{D}_y$ and generating a permutation distribution of (1): the p-value is equal to the share of permutation replicates (here, 1000) larger than the original value of the partial correlation coefficient.

When working with putative species that are isolated by distance, the within-group geographic dissimilarities (in $\mathbf{D}_x^{11}$ and $\mathbf{D}_x^{22}$) are usually small with respect to the between-group ones (in $\mathbf{D}_x^{21}$): by systematically altering the distribution of geographic distances within the groups, the permutation strategy might have a strong effect on the estimation of the partial correlation coefficient. A way to prevent this potential distortion is to assess significance in PMT using jackknife [4]. To do this, so-called pseudovalues for the partial correlation coefficient, $r_i^{\psi} = n \cdot r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x) - (n-1) \cdot r(\mathbf{D}_{y(i)}, \mathbf{D}_{g(i)}|\mathbf{D}_{x(i)})$, with $i = 1, \ldots, n$, are computed, where

$$r(\mathbf{D}_{y(i)}, \mathbf{D}_{g(i)}|\mathbf{D}_{x(i)}) = \frac{r(\mathbf{D}_{y(i)}, \mathbf{D}_{g(i)}) - r(\mathbf{D}_{y(i)}, \mathbf{D}_{x(i)})r(\mathbf{D}_{g(i)}, \mathbf{D}_{x(i)})}{\sqrt{(1 - r(\mathbf{D}_{y(i)}, \mathbf{D}_{x(i)})^2)(1 - r(\mathbf{D}_{g(i)}, \mathbf{D}_{x(i)})^2)}} \tag{2}$$

is the i^{th} jackknife replicate of the partial correlation coefficient and $\mathbf{D}_{(i)}$ is a dissimilarity matrix in which the row and corresponding column involving the i^{th}

individual in the dataset are excluded. The computation of these n pseudovalues, which are then used to carry out a one-sided t-test, takes into account the dependence in the data since all the dissimilarities involving the i^{th} individual are removed at each iteration.

Another option for assessing significance is nonparametric bootstrap, where individuals are resampled with replacement to yield B replicates (here, 1000) of the partial correlation coefficient, $r_b(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$, with $b = 1, \ldots, B$. Let $F_r(d) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(r_b(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x) \leq d\right)$ be the empirical distribution function of the bootstrap replicates. The lower boundary r_{lo} of bias-corrected bootstrap confidence intervals [4] is picked from the distribution of bootstrap replicates according to how far the original value of the partial correlation falls from its median bootstrap replicate. The null hypothesis of the PMT is rejected if $r_{lo} = F_r^{-1}\left(\Phi(2 \cdot \hat{\zeta} - \Phi^{-1}(0.95))\right) > 0$, where $\hat{\zeta} = \Phi^{-1}\left(F_r(r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x))\right)$ and Φ is the cumulative standard normal distribution function.

4 Simulations

`GSpace` [10] can simulate genetic data on a lattice under isolation by distance. Given the location of the sampled individuals on the lattice, the history of their neutral genes is simulated going backward through generations. At each generation, migration is handled by means of two-dimensional dispersal distributions, such as the uniform (for absence of IBD behavior) or the Zeta (for strong IBD behavior). Details on this simulation study can be found in [3].

For type-I-error rate evaluations, the two separate groups were generated with identical parameter settings and during the same software execution, in order for their individuals to have ancestors in common. For power evaluations, the two groups were simulated in two distinct software executions, still with the same parameter settings (e.g., IBD behaviour), but without any genetic information in common (genetic distance always equal to its maximum). These scenarios were investigated across a range of sample sizes and number of loci, so to assess the impact of data availability on the methods' performance. For each of these setups, 100 repetitions were simulated and the share of rejections from each method was recorded.

5 Results

The empirical type-I-error rate and power of the three versions of the PMT described above are displayed in Fig. 1, where they are arranged according to the IBD behaviour shared by the two groups in the simulation, the number of loci available for the computation of the shared allele distance and the number of individuals in each group. The most noticeable difference between the methods is the very large type-I-error rate often shown by the bootstrap-based PMT, which is due - as expected - to resampling with replacement: when identical individuals are resampled, this generates null distances that are outlying with respect to the

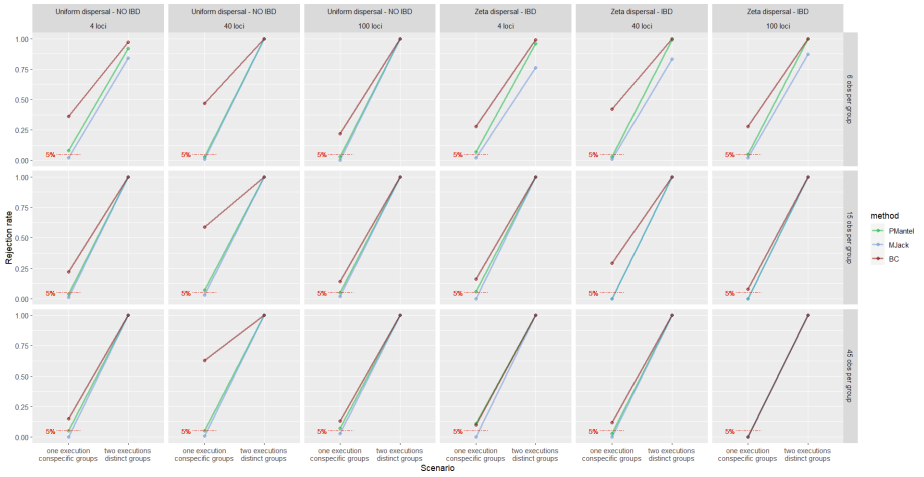


Fig. 1. Power plot based on `GSpace` data showing the empirical rejection rate over 100 replicates for each scenario. Faceted by a combination of IBD behaviour, number of loci and sample size. Scenarios under the null hypothesis are labelled "one execution, conspecific groups", while those under the alternative are labelled "two executions, distinct groups". A red dashed line is drawn at the significance level adopted in this study. (Color figure online)

bulk of the distance-distance data, leading to a biased distribution of bootstrap replicates of the partial correlation coefficient. Indeed, accounting for geographic separation, in those situations there appears to be a large positive association between genetic and grouping distance, as null distances can only be found when the grouping indicator is equal to zero. Consequently, bootstrap replicates have an upward bias and r_{lo} often falls above the zero threshold, leading to a rejection of the null hypothesis.

As far as the other two methods are concerned, the jackknife-based tests tend to be more conservative, with type-I-error rates very close to zero, whereas the permutation-based ones are often above the significance level, but seldom significantly so. Indeed, only in one scenario (Zeta dispersal, four loci, 45 individuals per group) permutation-based PMT has a rejection rate larger than 9%, i.e., a rate that is significant according to a binomial test with significance level equal to 5% in which each of the one-hundred tests is assumed to be distributed as a Bernoulli random variable with 5% success probability. On the other hand, in scenarios under the alternative hypothesis where any difference can be spotted (last three plots on first row), permutation-based tests are more powerful. In those setups, the tendency of within-group genetic distances to grow with geographic separation (IBD behaviour) sometimes implies that the jackknife-based PMT does not reject the null hypothesis of conspecificity, while permutation-based PMT looks less sensible to this.

6 Conclusions

In this study, three versions of the partial Mantel test - two of which never used for species delimitation before - were compared based on `GSpace` simulations. Bootstrap-based PMTs displayed very large type-I-error rates, probably because resampling with replacement inflated the number of null distances in the dataset. Jackknife-based PMTs were more conservative than permutation-based ones, but less powerful. Further investigation is required to explain the permutation-based PMTs' type-I-error rates sometimes above the significance level: these tests might be systematically too liberal as a result of the distortion the permutations introduce by altering the distribution of within-group and between-group geographic distances, something that the jackknife procedure does not imply.

The comprehensive study in [3] includes additional data scenarios (also from other simulators) and the investigation of other distance-distance methods for species delimitation, such as a newly introduced extension of the linear mixed model in [2] and the jackknife-based protocol proposed by [7]. These linear models can also accommodate different spread capacities of the analyzed species.

References

1. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L.: High resolution of human evolutionary trees with polymorphic microsatellites. Nature (London) **368**(6470), 455–457 (1994)
2. Clarke, R.T., Rothery, P., Raybould, A.F.: Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. J. Agric. Biol. Environ. Stat. **7**(3), 361–372 (2002)
3. d'Angella, G., Hennig, C.: Biological species delimitation based on genetic and spatial dissimilarity: a comparative study (2024). https://arxiv.org/abs/2401.12126
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York London (1993)
5. Frantz, A.C., Cellina, S., Krier, A., Schley, L., Burke, T.: Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? J. Appl. Ecol. **46**(2), 493–505 (2009). https://doi.org/10.1111/j.1365-2664.2008.01606.x
6. Hausdorf, B., Hennig, C.: Package 'prabclus', version 2.3-2 (2019). https://CRAN.R-project.org/package=prabclus
7. Hausdorf, B., Hennig, C.: Species delimitation and geography. Mol. Ecol. Resour. **20**(4), 950–960 (2020). https://doi.org/10.1111/1755-0998.13184
8. McRae, B.H.: Isolation by resistance. Evolution **60**(8), 1551–1561 (2006). http://www.jstor.org/stable/4095372
9. Medrano, M., López-Perea, E., Herrera, C.M.: Population genetics methods applied to a species delimitation problem: Endemic trumpet daffodils (narcissus section pseudonarcissi) from the southern iberian peninsula. Int. J. Plant Sci. **175**(5), 501–517 (2014)
10. Virgoulay, T., Rousset, F., Leblois, R.: Gspace: an exact coalescence simulator of recombining genomes under isolation by distance. Bioinformatics **37**(20), 3673–3675 (2021)



Smooth Logistic Mass Univariate Inference for MS Lesion Data Using Sign-Flipping

Samuel Davenport[1(✉)], Riccardo de Santis[2], Jesse Hemerik[3], Jelle Goeman[4], and Livio Finos[5]

[1] Division of Biostatistics, University of California San Diego, San Diego, USA
sdavenport@health.ucsd.edu
[2] Riccardo De Santis Department of Economics and Statistics, University of Siena, Siena, Italy
riccardo.desantis2@unisi.it
[3] Econometric Institute, Erasmus University, Rotterdam, The Netherlands
hemerik@ese.eur.nl
[4] Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
j.j.goeman@lumc.nl
[5] Department of Statistical Sciences, University of Padova, Padua, Italy
livio.finos@unipd.it

Abstract. Magnetic resonance imaging (MRI) has become a key technique in the study of multiple sclerosis. In a typical analysis pipeline binary lesion maps are typically modelled voxelwise using either standard linear models or logistic regression, in order to detect relationships with variables of interest. Fitting linear models does not respect the binary nature of the data and, as we shall show, this can lead to false positives. Instead voxelwise logistic regression correctly models the data however it suffers from a lack of power. Smoothing is a key means of increasing the signal to noise ratio in neuroimaging data but has previously been thought not to be applicable to lesion based logistic regression as, when applied to the 0/1 maps, it destroys the binary nature of the data [9]. In this work we show how to combine smoothing with logistic regression modelling: increasing the power of inference whilst maintaining validity. In particular we develop a test based on sign-flipping the smooth effective score contributions and control for multiple testing using the sign-flipped distribution of the maximum smoothed score test-statistic.

Keywords: Multiple testing · Generalized Linear models · Lesion Data · Sign flipping

1 Introduction

In mass-univariate modelling of 0/1 lesion based data the commonly used approach of fitting linear models to the binary data (c.f. [3,12]) can lead to false positives as such models are evidently misspecified. Modelling outcomes restricted

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 525–531, 2025.
https://doi.org/10.1007/978-3-031-64431-3_87

to 0/1 as if they were continuous does not properly acknowledge the mean-variance relationship inherent in binary random variables, and may easily result in predicted values outside the 0/1 range, especially if covariates are present. As a result of this model misspecification, type I error control is not guaranteed. Better specified logistic models and their corresponding tests correctly control the error rate but in the presence of multiple tests suffer from a lack of power.

Recently, permutation-type methods have been developed that enable inference in general GLMs [7]. These methods can be combined with permutation-based multiple testing which naturally takes into account the dependence structure between the tested variables [14]. In particular, in the present setting where many binary responses are considered, this approach allows for multiple testing of associations between the many binary responses and the predictor of interest. Since the binary responses are highly correlated, using such permutation-based multiple testing methods leads to large gains in power compared to simply computing an individual p-value for every voxel and using Bonferroni-Holm to correct for multiple comparisons.

When analyzing imaging data, smoothing is a commonly used pre-processing step, which is applied to increase the signal to noise ratio. This makes it easier to detect effects of interest. Smoothing has been combined with fitting linear models to lesion data [3] where the 0/1 maps are smoothed themselves. As remarked in [9] this approach destroys the binary nature of the data and does not guarantee control of the error rate. We propose to fit GLMs at each voxel and to instead smooth the resulting score contributions. We shall demonstrate that this smoothing step can be combined with sign-flipping, resulting in a test which respects the model structure while increasing power and controlling false positive rates.

Matlab code implementing our methods is available in the matPerm toolbox [4]. Python implementations are also available in the pyperm python package [5]. Smoothing and vizualization were performed via the StatBrainz matlab toolbox [6].

2 Methods

2.1 Setting

Suppose, given a number of subjects $n > 0$, that we observe random i.i.d. binary images $Y_1, \ldots, Y_n$ on a set of voxels $\mathcal{L} \subset \mathbb{R}^3$, i.e. such that $Y_i(l) \in \{0, 1\}$ for all $l \in \mathcal{L}$ for each $1 \leq i \leq n$. Moreover, given $p > 0$, suppose that $x_i \in \mathbb{R}^p$ is a vector containing the covariates of interest for the ith subject.

In order to account for the binary nature of the data we shall model this data by fitting a generalized linear model with a logit link function. To do so at each $l \in \mathcal{L}$ we shall assume that $Y_i(l) \sim \text{Binomial}(q_i(l))$ where $q_i : \mathcal{L} \to \mathbb{R}$ is a map giving the chance of a lesion occurring at each voxel for the ith subject. Assume that for each $l \in \mathcal{L}$,

$$\log\left(\frac{q_i(l)}{1 - q_i(l)}\right) = x_i^T \beta(l) + z_i^T \gamma(l) \tag{1}$$

where $(x_i)_{i=1}^n \in \mathbb{R}^p$ and $(z_i)_{i=1}^n \in \mathbb{R}^r$ are vectors consisting of the interest and nuisance covariates respectively for each subject and $\beta(l) \in \mathbb{R}^p, \gamma(l) \in \mathbb{R}^r$. We will also write $X = [x_1, \ldots, x_n]^T$ and $Z = [z_1, \ldots, z_n]^T$ and $Y(l) = [Y_1(l), \ldots, Y_n(l)]^T$.

At each voxel $l \in \mathcal{L}$, we will want to test the null hypothesis

$$H_0(l) : \beta(l) = 0.$$

Doing so introduces a multiple testing problem over voxels which can be severe. Let $\mathcal{N} = \{l \in \mathcal{L} : \beta(l) = 0\}$ be the set of voxels at which the null hypothesis holds.

2.2 Obtaining and Smoothing the Effective Score Contributions

For each voxel, the test statistic that we use is the *effective score* [13], which is the sum of n terms which we call the score contributions. These score contributions have mean 0 under H_0 [10] and so can be used in order to test the null hypothesis. The effective score at a voxel l can be written as

$$S(l) = n^{-1/2} X^T W(l)^{1/2} (I - P(l)) W(l)^{-1/2} (Y(l) - \hat{\mu}(l)).$$

Here $\hat{\mu}(l)$ is a vector giving the fitted values and $W(l)$ is a diagonal matrix with the GLM weights - these are equal to the variances as the link function is logistic. Moreover, following [7], $P(l)$ is the GLM projection matrix defined as

$$P(l) = W(l)^{1/2} Z \left(Z^T W(l) Z\right)^{-1} Z^T W(l)^{1/2}.$$

$S(l)$ is the inner product of two n-vectors, so that it is the sum of n terms:

$$S(l) = n^{-1/2} \sum_{i=1}^{n} \left[X^T W(l)^{1/2} (I - P(l)) W(l)^{-1/2}\right]_i (Y_i(l) - \hat{\mu}_i(l)) =: n^{-1/2} \sum_{i=1}^{n} \nu_i(l).$$

The summands $\nu_1(l), ..., \nu_n(l)$ are called the effective score contributions. Given a smoothing kernel $K : \mathbb{R}^3 \to \mathbb{R}$ with support $\text{supp}(K) \subset \mathbb{R}^3$, we obtain the smoothed score contribution for each subject i as the image $\tilde{\nu}_i : \mathcal{L} \to \mathbb{R}$ such that

$$\tilde{\nu}_i(l) = \sum_{l' \in \mathcal{L}} K(l - l') \nu_i(l).$$

Let $\hat{\mu}_n : \mathcal{L} \to \mathbb{R}$ be the mean, computed pointwise, of the smoothed score contributions. I.e. for all $l \in \mathcal{L}$, define $\hat{\mu}_n(l) = \frac{1}{n} \sum_{i=1}^n \tilde{\nu}_i(l)$. A score based test-statistic can then be constructed as $T_n : \mathcal{L} \to \mathbb{R}$, such that for each $l \in \mathcal{L}$,

$$T_n(l) = \sqrt{n} \hat{\mu}_n(l).$$

As $n \to \infty$, $(T_n(l) : l \in \mathcal{L})$ converges to a multivariate Gaussian distribution, with limiting variance $G = \lim_{n \to \infty} \text{var}(\tilde{\nu}_1)$. This distribution can be targeted using sign-flipping as we demonstrate in the next section.

2.3 Performing Inference by Sign-Flipping

In order to control the FWER over voxels we shall target the distribution of $\max_{l\in\mathcal{L}} T_n(l)$ using sign-flipping based resampling. In particular we shall sign-flip the smooth effective score, allowing us to combine smoothing with the GLM resampling and multiple testing inference approaches of [11] and [14]. To do so, given a number of sign-flips $B \in \mathbb{N}$, let

$$T_n^b(l) = n^{-1/2} \sum_{i=1}^{n} g_{bi} \tilde{v}_i(l),$$

where we take $g_1 = (1, \ldots, 1) \in \mathbb{R}^n$ to preserve the observed test statistic, and for $2 \leq b \leq B$ define $g_b = (g_{b1}, \ldots, g_{bn})$ to be independent and uniformly distributed on $\{-1, 1\}^n$. Using the same sign flip at each voxel allows the spatial dependence structure to be preserved within each flip. Asymptotically the sign-flipped test-statistics converge to the same limiting distribution as the original test-statistics.

Theorem 1. *For each* $2 \leq b \leq B$, $vec(T_n^b) \overset{d}{\Longrightarrow} N(0, G)$.

Thus, in order to target the distribution of the maximum for α in $(0, 1)$, we can apply the max-T approach of [16] and let

$$\lambda^* = \inf \left\{ \lambda : \frac{1}{B} \sum_{b=1}^{B} 1[\max_{l\in\mathcal{L}} T_n^b(l) \geq \lambda] \geq 1 - \alpha \right\}.$$

Then consider the multiple testing procedure $\mathcal{M}$ which rejects $H_0(l)$ at the voxels l such that $T_n(l) > \lambda^*$ and let $\mathcal{R}_n$ be the set of rejected voxels. Theorem 2 shows that, asymptotically, $\mathcal{M}$ has strong control of the familywise error rate up to the support of the kernel (c.f. [15]).

Theorem 2. *Let* $\mathcal{N}(K) := \{l \in \mathcal{L} : \nexists l' \in \mathcal{N}^C \text{ with } l - l' \in supp(K)\}$. *Then*

$$\lim_{n\to\infty} \mathbb{P}\left(|\mathcal{R}_n \cap \mathcal{N}(K)| > 0\right) \leq \alpha.$$

3 Results

3.1 False Positive Simulations

In order to demonstrate the need to apply well-specified models we consider a simple case of (1). In particular, letting $n = 200$, $p = 1$ and $r = 3$, we take $(z_{i2}, z_{i3})_{i=1}^n \overset{i.i.d.}{\sim} N(0, I_2)$, where I_2 is the 2×2 identity matrix, and take $z_{i1} = 1$ and $x_{i1} = z_{i1} z_{i2}$ for all $1 \leq i \leq n$. We will set $\gamma = [-5, 1, 5]$, let $\beta = 0$ and test for the presence of the interaction effect x. We generate 1000 realisations of this model and compute p-values using the effective score test baed on sign-flipping, the Wald test and the linear model. The empirical CDF of each over the realizations is plotted in Fig. 1.

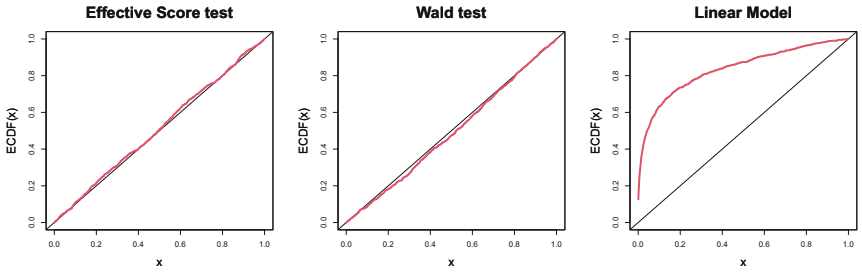


Fig. 1. Empircal CDF of the simulated p-values. Fitting a linear model to the data results in high levels of false positives. Instead the sign-flipped effective score test and Wald test control to the nominal rate.

As can be seen from the results the effective score test and the Wald test control the false positive rate to the nominal level. Instead fitting a linear model to this data can lead to large levels of false positives. The advantage of the effective score over the Wald test, as we see below, is that it can account for the correlation structure when adjusting for multiple comparisons. This allows for more powerful inference.

3.2 Fitting the Models to the MS Lesion Data

We have lesion data from 238 subjects with multiple sclerosis. After processing, around 40,000 lesions had at least one lesion in at least one subject. For each subject we have covariates for age, sex, EDSS, PASAT and an indicator of the MS disease progression divided into 2 subtypes: mild and severe. We run logistic regression at each voxel with these covariates and an intercept and are interested in testing for the effect of PASAT. We calculate the effective score contributions and smooth these as described in Sect. 2.2, using a 3D isotropic Gaussian Kernel with a full width at half maximum of 4 voxels, truncated at 3 voxels. We rejected the null hypothesis using the testing procedure $\mathcal{M}$ described in Sect. 2.3. The results are displayed in Fig. 2. They show an effect of PASAT on lesion incidence in both the left and right hemispheres with a greater area of activation in the right hemisphere.

Instead, analyzing this data using the Wald test, calculating p-values using Firth regression and correcting for multiple comparisons using Bonferroni-Holm, no significant voxels remain. Nor are any signficant voxels are detected even if the more liberal Benjamini-Hochberg algorithm [1] is applied. These data were also analysed in [9] with Firth regression, however their results were not corrected for multiple comparisons. Instead they thresholded the Wald test-statistics at ± 2. Since the data consists of a large number of voxels correction for multiple comparisons is both necessary and important, as is widely recognized in neuroimaging [2,8,17]. With our approach we are thus more sensitive to the presence of effects of interest even after controlling for multiple comparisons.

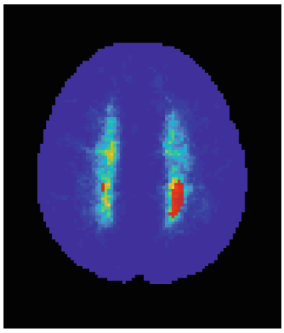


Fig. 2. A slice through the results of applying our approach at the level 5%, significant voxels are shown in red. The background displays the lesion density over the 238 subjects.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc.: Ser. B (Methodol.) **57**(1), 289–300 (1995)
2. Bennett, C.M., Miller, M.B., Wolford, G.L.: Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. Neuroimage **47**(Suppl 1), S125 (2009)
3. Charil, A.: Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: application to 452 patient data sets. Neuroimage **19**(3), 532–544 (2003)
4. Samuel Davenport. matPerm matlab toolbox. https://github.com/sjdavenport/matPerm (2024)
5. Samuel Davenport. pyperm python package. https://github.com/sjdavenport/pyperm (2024)
6. Samuel Davenport. StatBrainz matlab toolbox. https://github.com/sjdavenport/StatBrainz (2024)
7. De Santis, R., Goeman ,J.J., Hemerik, J., Davenport, S., Finos, L.: Inference in generalized linear models with robustness to misspecified variances. arXiv preprint http://arxiv.org/abs/2209.13918arXiv:2209.13918 (2022)
8. Eklund, A., Nichols, T.E., Knutsson, H.: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proce. Nat. Acad. Sci. **113**(28), 7900–7905 (2016)
9. Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T.E., Johnson, T.D.: Analysis of multiple sclerosis lesions via spatially varying coefficients. Ann. Appl. Stat. **8**(2), 1095 (2014)
10. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. Royal Stat. Soc. Series B (Statistical Methodology) **82**(3), 841–864 (2020)
11. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. Royal Stat. Soc. Series B-stat. Method. **82**(3), 841–864 (2020)

12. Lampe, L.: Visceral obesity relates to deep white matter hyperintensities via inflammation. Ann. Neurol. **85**(2), 194–203 (2019)
13. Marohn, F.: A comment on locally most powerful tests in the presence of nuisance parameters. Commun. Stat. Theory Meth. **31**(3), 337–349 (2002)
14. De Santis, R., Goeman, J.J., Davenport, S., Hemerik, J., Finos, L.: Permutation-based multiple testing when fitting many generalized linear models. arXiv preprint http://arxiv.org/abs/2403.02065arXiv:2403.02065 (2024)
15. Telschow, F., Davenport, S.: Riding the SuRF to continuous land: precise FWER control for gaussian random fields. Preprint (2023)
16. Westfall, P.H., Young, S.S.: Resampling-based multiple testing: Examples and methods for p-value adjustment. Wiley (1993)
17. Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P.: A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. **12**(6), 900–918 (1992)



Penalised Spatial Quantile Regression: Application to Air Quality Data

Marco F. De Sanctis[1], Ilenia Di Battista[1(✉)], Eleonora Arnone[2], Cristian Castiglione[3], Mauro Bernardi[3], Alessandro Palummo[1], and Laura M. Sangalli[1]

[1] MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo Da, Vinci 32, Milano 20133, Italy
{marcofrancesco.desanctis,ilenia.dibattista, alessandro.palummo}@mail.polimi.it, 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

[3] University of Padova, Department of Statistical Sciences, Via Cesare Battisti, 241, Padova 35121, Italy
{cristian.castiglione,mauro.bernardi}@unipd.it

Abstract. This article explores the application of quantile regression techniques to capture non-standard tail behaviours in spatially correlated data, typically encountered in environmental and climate sciences. In particular, we propose extensions of penalised spatial quantile regression models, to accommodate spatio-temporal data, as well as simultaneous estimates of spatial quantile surfaces. Through a real data application in the Lombardy region, we demonstrate the efficacy of the proposed models in analysing measurements of NO_2 concentrations, showcasing the utility of quantile regression, where the spatial mean provides poor or little information on the phenomenon under study.

Keywords: quantile regression · environmental science · spatio-temporal data · simultaneous quantile estimation

1 Introduction

The development of statistical models for complex spatial phenomena is crucial in many fields, like environmental and climate sciences. Despite its importance, traditional approaches to analyse mean surfaces often fall short in capturing non-standard tail behaviours exhibited by the data distribution. These deviations, including heteroscedasticity and localised skewness, are commonly encountered in the analysis of spatially correlated data [6]. Quantile regression allows for the estimation of conditional quantiles of a variable of interest, providing insights into the entire distribution rather than just the mean.

M. F. De Sanctis and I. Di Battista: These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 532–537, 2025.
https://doi.org/10.1007/978-3-031-64431-3_88

In this work, we analyse the NO_2 concentration measurements recorded in the 84 control units of the Agenzia Regionale per la Protezione dell'Ambiente (ARPA), located in the Lombardy region during December 2018, as illustrated in Fig. 1. Such environmental data are typically marked by strong skewness, indicating a substantial deviation from the Gaussian assumption of the errors. Furthermore, air pollutant data collected in Lombardy exhibit complex behaviours, due to the complex conformation of the territory, bordered at north by mountains and encompassing the Po valley, as well as localised extreme values, particularly in proximity to large urban centers. These outliers are of particular interest to scientists monitoring health conditions of the local population. Indeed, the long-term exposure to elevated levels of NO_2 concentrations poses various health risks to the residents, including allergic responses, respiratory symptoms, decreased lung function growth in children, cardiopulmonary effects, and even premature death [4].

In this work, we start from the model for spatial quantile regression proposed by [6]. Such method permits to model strong anisotropies and non-stationarities in the data, that may for instance derive from external forces, such as winds or currents. Figure 1 shows for instance the winds observed on the territory, on the date corresponding to the data represented in the same figure.

We extend the model presented in [6] toward two important directions. First of all, we generalize the model in order to encompass spatio-temporal data. Secondly, we generalize it toward simultaneous estimation of multiple spatial quantile surfaces, with non-crossing constraints. The augmentation towards space-time data allows to integrate temporal information with spatial data, enhancing the applicability of the resulting model. Indeed, incorporating temporal trends can unveil crucial dynamics influencing the spatial distribution of phenomena, fostering a deeper comprehension of the underlying processes. The generalization toward multiple quantiles, instead, ensures the monotonicity of the quantile estimates, promoting consistency and leading to more interpretable results. Additionally, a model ensuring non-crossing quantiles yields to more robust and accurate predictions [5].

2 Mathematical Model

Let $\{\mathbf{p}_i\}_{i=1..n}$ be n spatial locations over a domain $\Omega \subset \mathbb{R}^2$, with regular boundary $\partial\Omega$, and consider $\{Y_i\}_{i=1..n}$ real independent random variables with absolutely continuous distribution gathered on the spatial locations $\mathbf{p}_1, ..., \mathbf{p}_n$. Let $Q_{Y_i|\mathbf{x}_i,\mathbf{p}_i}(\alpha)$ be the conditional quantile function of Y_i for any probability level $\alpha \in (0, 1)$. The semiparametric model for the $\alpha-$quantile reads

$$Q_{Y_i|\mathbf{x}_i,\mathbf{p}_i}(\alpha) = \mathbf{x}_i^T\boldsymbol{\beta} + f(\mathbf{p}_i), \quad i = 1, ..., n \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^q$ is the vector of covariates observed at location $\mathbf{p}_i$, $\boldsymbol{\beta}$ is the regression vector, and $f : \Omega \to \mathbb{R}$ is a nonparametric term, defined as a spatial field over the domain Ω. The nonparametric term f permits a highly flexible modeling of the space and space-time variation of the phenomenon under study,

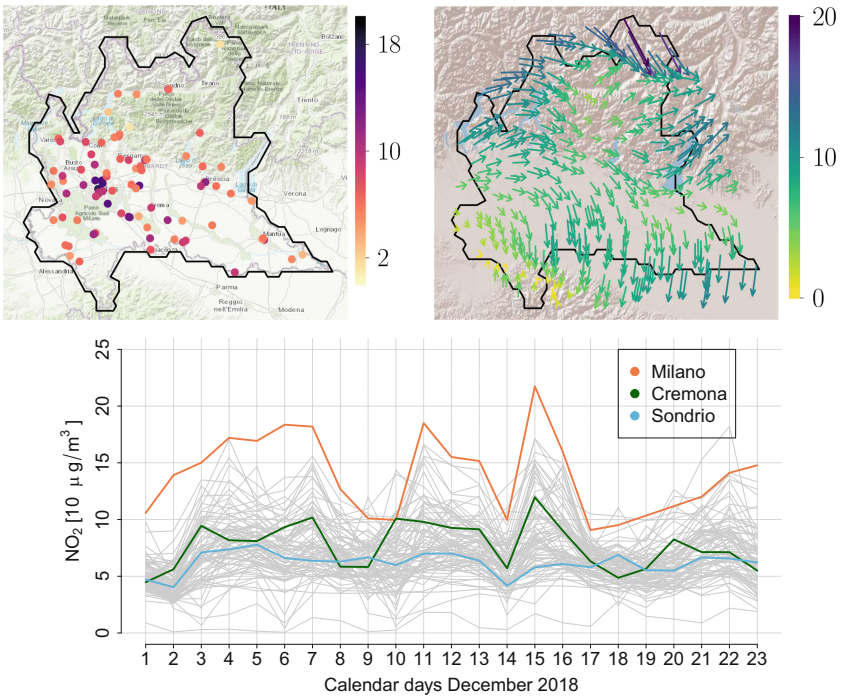


Fig. 1. Top left: daily maximum NO_2 concentration at the 84 ARPA control units in Lombardy, on 11^{th} December 2018. Top right: wind speed and direction in Lombardy, on 11^{th} December 2018. The intensities are proportional to the arrows' length. Bottom: temporal evolution of NO_2 at ARPA control units over three weeks, from 1^{st} to 23^{th} December. We highlight 3 profiles corresponding to 3 control units situated in the city of Milano (light brown), Cremona (green) and Sondrio (light blue).

as reviewed in linear regression settings by, e.g., [9]. The goal is to estimate f and $\boldsymbol{\beta}$ to reconstruct the quantile field Q for any level α. This can be done by minimizing a penalised loss functional $L+\lambda R$, where the loss L is an appropriate data fidelity criterion, and the regularizing term R accounts for the smoothness of the nonparametric part of the field. In particular, we consider the loss $L = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - f(\mathbf{p}_i))$, where the function $\rho_\alpha(x) := \frac{1}{2}|x| + (\alpha - \frac{1}{2})x$, known as quantile check function, is exploited to center the estimate on the correct quantile location [8]. The regularising term R involves a second order elliptic operator, applied to the spatial field f, enabling the inclusion of physics-informed problem-specific information [9]. In this way, it is for instance possible to embed into the estimation problem a space-varying anisotropic diffusion term, that model non-stationary anisotropy in the data, as well as a space-varying transport effect producing a smoothing towards a unique direction, that may for instance be the effect of winds. Such regularising term can be generalised into the spatio-temporal settings, using regularizing terms alike those considered in [1,3].

This permits to extend the semiparametric spatial quantile regression model in (1) to space-time settings. It should be noted that the estimation problem does not admit an analytical solution, therefore one must rely on appropriate iterative algorithms. Moreover, the infinite dimensional problem can be approximated via appropriate basis expansions, as implemented in the models reviewed by [9].

3 Application to Air Quality Data

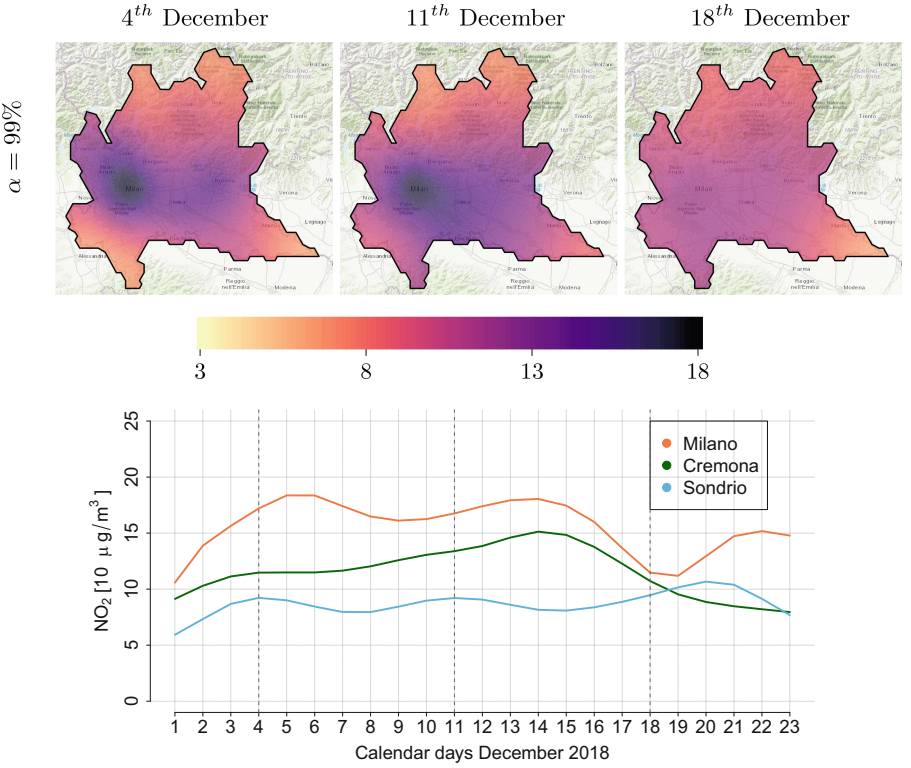


Fig. 2. Top: Estimated NO_2 99% space-time quantile surfaces, plotted at three time instances over Lombardy. Bottom: Time trend of the estimated 99% quantile of NO_2 concentration at three fixed spatial locations, in the cities of Milano (light brown), Cremona (green) and Sondrio (light blue).

Here we briefly show an application of proposed models to the NO_2 concentration measurements in Lombardy in December 2018.

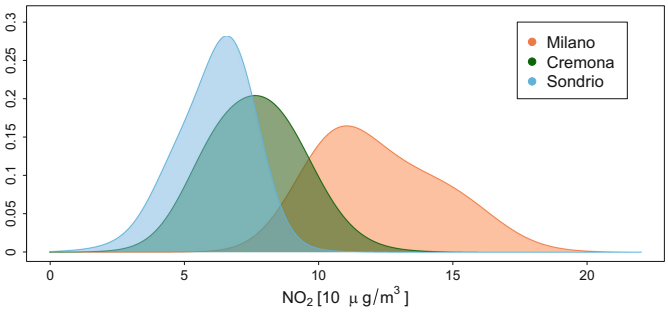


Fig. 3. Estimate of NO_2 distribution at Milano (light brown), Cremona (green) and Sondrio (light blue), obtained from the monotone quantile estimated fields.

Figure 2 shows the space-time estimate of the 0.99 quantile of NO_2 concentration in the whole territory, and the time trends of the 0.99 quantile in the same locations evidenced in the bottom panel of Fig. 1, namely Milano, Cremona and Sondrio. Figure 2 highlights the severity of pollution in the major urban centers, in particular in Milano and Brescia. Looking at the time trends in the bottom panel of Fig. 2, we can note that the concentration in Milano (light brown) exhibits a weekly pattern, with typically lower values on the weekends (1–2, 8–9, 15–16 and 22–23 December), coherent with people moving away from the city during weekends. In particular, a significant decrease in NO_2 concentration is appreciable in corresponds with the bank holiday of 7–8-9 December, with a reduction in the pollutant which benefit also to subsequent days, and a further strong decrease in correspondence of heavy rains during 17–18 December.

Moving to the context of simultaneous estimates of non-crossing quantiles, we consider the NO_2 measurements on a fixed day, specifically the 11^{th} December. We consider as covariates population density and altitude at the observed locations. Moreover, we consider a regularizing term that includes a transport effect derived from wind data recorded on the considered day. In particular, the weight of the transport term is optimized using a two step procedure, as proposed by [2]. As expected, the model fit returns a positive effect for population density, and a negative effect for altitude. Such effects are proved to be significant on the mean, as tested by appropriate Wald tests, see, e.g., [7]. Owing to the monotonicity of the estimated quantile surfaces, it is possible to reconstruct the density profile of the NO_2 concentration at fixed spatial points. Figure 3 shows the probability density function in the cities of Milano, Cremona and Sondrio. We can observe that the density supports and modes shift towards higher values, moving from less polluted areas, as the mountain town of Sondrio, to more critical ones, like the metropolitan city of Milano. In particular, the density in Milano is characterized by a heavy right tail, which exhibits the harmful situation faced by the citizens.

4 Discussion

A challenging direction of future research concerns the development of appropriate inference procedures for the proposed spatial quantile regression models, along the ideas discussed in [7] for linear regression settings.

Acknowledgements. Research funded by 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

1. Arnone, E., Sangalli, L.M., Vicini, A.: Smoothing spatio-temporal data with complex missing data patterns. Stat. Model. **23**(4), 327–356 (2023)
2. Bernardi, M.S., Carey, M., Ramsay, J.O., Sangalli, L.M.: Modeling spatial anisotropy via regression with partial differential regularization. J. Multivar. Anal. **167**, 15–30 (2018)
3. 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. Stoch. Env. Res. Risk Assess. **31**(1), 23–38 (2017)
4. Board, C.A.S.: California air sources board. https://ww2.arb.ca.gov/resources/nitrogen-dioxide-and-health
5. Bondell, H.D., Reich, B.J., Wang, H.: Noncrossing quantile regression curve estimation. Biometrika **97**(4), 825–838 (2010)
6. Castiglione, C., et al.: Penalized quantile regression for spatial distributed data. In: Proceedings of the GRASPA 2023 Conference, pp. 1–6 (2023)
7. Ferraccioli, F., Sangalli, L.M., Finos, L.: Nonparametric tests for semiparametric regression models. TEST **32**(3), 1106–1130 (2023)
8. Koenker, R., Ng, P., Portnoy, S.: Quantile smoothing splines. Biometrika **81**(4), 673–680 (1994)
9. Sangalli, L.M.: Spatial regression with partial differential equation regularisation. Int. Stat. Rev. **89**(3), 505–531 (2021)

Sign-Flip Test for the Cox Regression Model

Riccardo De Santis[1(✉)], Jelle J. Goeman[2], Hein Putter[2], and Livio Finos[3]

[1] Department of Economics and Statistics, University of Siena, Siena, Italy
riccardo.desantis2@unisi.it
[2] Biomedical Data Sciences Leiden University Medical Center, Leiden, The Netherlands
j.j.geoman@lumc.nl, h.putter@lumc.nl
[3] Department of Statistical Sciences, University of Padova, Padua, Italy
livio.finos@unipd.it

Abstract. Cox regression model is a popular tool in survival analysis, whose aim is to quantify the impact of covariates on the survival times. The relevance of the coefficients is usually tested through a parametric test. However, the properties of this test are only asymptotical and can show a slow convergence to the nominal level. We propose a different approach to perform the test based on sign-flipping of the score contributions. We have defined two alternatives which show a faster convergence to the nominal level of the test.

Keywords: Cox regression model · Survival analysis · Score test · Sign flipping

1 Introduction

Time to event data are particular type of data which arise in many applied fields, such as medicine, biology and demography. Let T indicate the time, assumed to be continuous, at which the random variable associated to an event of interest, for example the death, is observed. Researchers are often not interested on the underlying distribution $f(t)$, but instead they focus on the hazard rate $\lambda(t)$. To be precise, we assume to have a sample of n observation, and for each unit i we observe (Y_i, δ_i), where $Y_i = \min\{T_i, C_i\}$ is the observed time, which might be T_i - the event is observed - or C_i - the individual left the study, while $\delta_i = 1\{T_i \leq C_i\}$ is the event indicator, taking the value of 1 if the event is observed. When some explanatory variables are observed we might be interested in studying the influence of them on the survival times. A widely used and popular model is the Cox regression model [3], which models the impact of the covariates on the hazard ratio and is defined as

$$\lambda(y_i|x_i, z_i) = \lambda_0(y_i) \exp\{\beta^T x_i + \gamma^T z_i\}.$$

The model is called semiparametric in the sense that the baseline hazard $\lambda_0(y_i)$ is left unspecified, while (X_i, Z_i) represents observed covariates and (β, γ) are regression parameters which have to be estimated. Hereafter, we assume that $\dim\{\beta\} = 1$ and $\dim\{\gamma\} = q$ which is treated as fixed. We will consider β as the parameter of interest, while γ is treated as a nuisance parameter.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 538–542, 2025.
https://doi.org/10.1007/978-3-031-64431-3_89

The model is generally fitted through a partial likelihood approach, defined as

$$L_p = \prod_{i \in D} \left\{ \frac{\exp\{\beta^T x_i + \gamma^T z_i\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right\}$$

where D represents the set of uncensored observations, while R_i represents the risk set at time i, which is the set of individuals who are still under study at a time just prior to Y_i. Note that this likelihood is a product over only the elements which are uncensored, equal to the cardinality of D. The use of the partial likelihood as a valid tool for inference was first justified by [4]. However, note that it can be seen also as a profile likelihood from the full-censored likelihood [8]. The maximum likelihood estimates of the parameters are obtained by maximizing the partial likelihood, which is therefore treated as a usual likelihood.

Our interest is on the regression parameters, considering the null hypothesis

$$H_0 : \beta = \beta_0,$$

against a one or two-sided alternative, while the other parameters are nuisance parameters, both the regression parameters γ and the baseline hazard $\lambda_0(y)$.

In absence of nuisance regression parameters, a large number of proposals have been done about two-sample or general k-sample problems, when the hypothesis is on the equality of the survival curves, considering different censoring schemes. These methods are either based on ranks, sign-flips or permutations, and reflect the nonparametric nature of the estimation of the survival function. A general overview is given by [1,9]. When testing for the effect of a continuous variable in the Cox model, without nuisance covariates, a permutation-based solution is given by [12].

In presence of nuisance regression parameters γ general nonparametric solutions are not available. Concerning the null hypothesis about the target parameter β, the parametric approach for this kind of test consists of three alternatives which have similar behavior, especially for growing sample size, respectively the Wald test, the Likelihood ratio test (LRT) and the parametric Score test. However, we will see in Sect. 3 that these tests show an anti-conservative behavior for small sample sizes and demands for an improvement.

2 Sign-Flip Tests for the Cox Model

Let ℓ be the log-likelihood of a parametric model, assumed to be a sum of q contributions, let ℓ_β, ℓ_γ be the first partial derivatives of the log-likelihood and $\ell_{\beta\beta}, \ell_{\beta\gamma}, \ell_{\gamma\gamma}$ be the second derivatives. For parametric models [6] define the effective score as

$$S = \ell_\gamma - \ell_{\beta\gamma}(\ell_{\gamma\gamma})^{-1}\ell_\gamma = \sum_{i=1}^{q} v_i.$$

The authors proposed to perform a test by sign-flipping of the summands underlying the effective score, that is, randomly multiplying by -1 and 1 the contributions that we have denoted as v_i. Call $S(F)$ the flipped test statistic, where F is the matrix representing the

flip. It can be defined as a diagonal matrix whose elements are -1 and 1. Note that for the identity matrix we recover the observed test statistic.

The test is thus performed by comparing the observed test statistic with the quantile of the conditional flipping distribution. The test is shown to be asymptotically exact [6]. A convergence improvement for Generalized Linear Models (GLMs) is obtained by [5]. They find an explicit formula for $Var\{S(F)\}$ and define the new test statistic as $S^*(F) = S(F)/Var\{S(F)\}$. The test is shown to be second-moment exact, whose simulation shows appreciable behavior for small sample size.

About the Cox regression model, a first intuitive proposal is based on sign-flipping of the contributions underlying the partial likelihood. First of all, remember that the partial likelihood is a profile likelihood with respect to the baseline hazard. Moreover, the effective score can be deduced from the profile likelihood [11]. Hence, we can apply directly the test of [6], where the total of the contributions to be flipped is equal to the number of uncensored observations.

A second proposal is based on a GLM formulation of the Cox model investigated by [7, 10]. The authors noted that the likelihood of the full-censored Cox model coincides with the likelihood of a particular Poisson regression model, where each original observation produces a new observation for each time event for which it is still surviving, while the model is fitted with a nuisance parameter per each time event. Hence we can directly apply the standardized test for GLMs; the main weakness of this approach is given by the computational burden required. Indeed, a moderate original sample size can easily turn on a massive sample size in the Poisson formulation.

3 Simulation Study

We perform a simulation study in order to evaluate the capability of controlling the type I error. We compare the three parametric methods - Wald test, LRT and parametric Score test - with the two Flipscores test, the one based on the full GLM formulation - Flip GLM - and the one based on the partial likelihood - Flip Partial lik. Note that for the GLM formulation we have used the standardized test - the most computational intensive procedure - while for the partial likelihood formulation we have used the effective test, that is, a less computational demanding procedure. A total of 5 000 simulations have been run. The data are generated through the inverse probability method [2]. The baseline hazard is a Weibull distribution with shape and scale parameters (under the complete null hypothesis) equal to 1. The censoring time are generated through an independent exponential distribution, whose parameter is chosen in order to control the average proportion of censored observations. The model contains two covariates (X, Z). In the Figs. 1-3 the left panel considers a correlation of 0 between X and Z, while such correlation is 0.3 in the right panel.

Figure 1 shows the type I error control when testing a continuous variable. The true β is set equal to 0 and we test $H_0 : \beta = 0$ with significance level $\alpha = 0.05$. The parameter γ is set equal to 0.5. The x axis represents the sample size, while the y axis represents the empirical type I error. The dotted lines represent a 95% simulation confidence bands. We see that, even for a sample size of 60, the parametric tests are anti-conservative, beyond the confidence bands. On the other hand, the two Flipscores tests show a similar and more appreciable behavior.

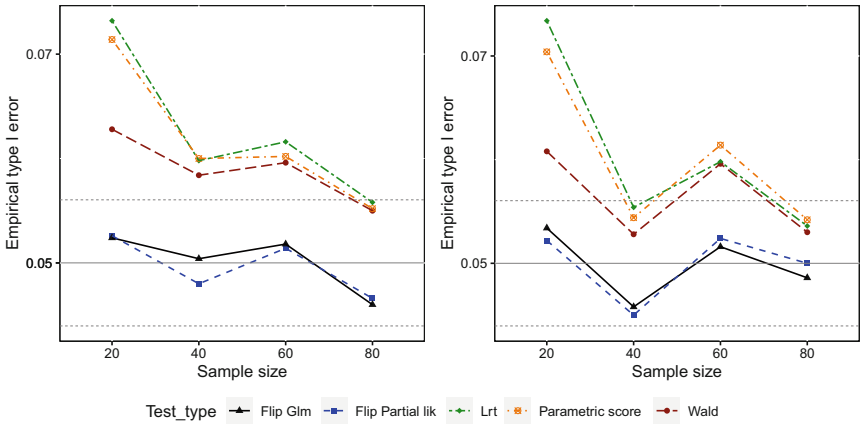


Fig. 1. Type I error control for continuous covariate. Average censoring equal to 25%.

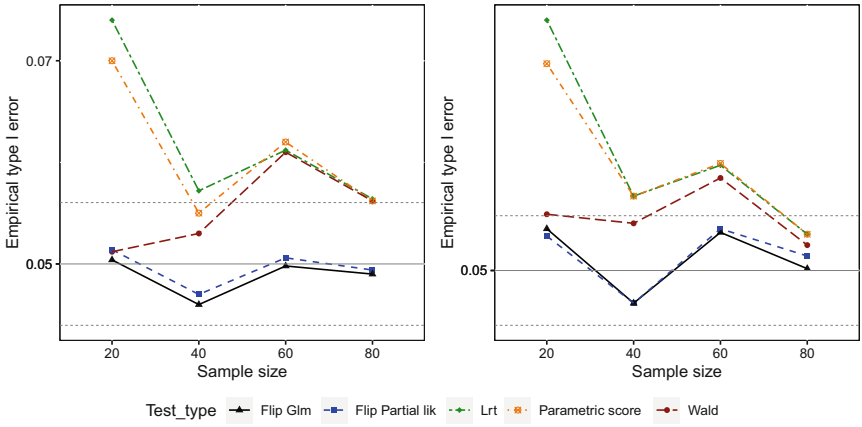


Fig. 2. Type I error control for continuous covariate. Average censoring equal to 50%.

Figure 2 shows the type I error control when testing a continuous variable, when the average proportion of censored observations is 50%. The true β is set equal to 0 and we test $H_0 : \beta = 0$ with significance level $\alpha = 0.05$. The parameter γ is set equal to 0.5. Despite the fact that the effective number of observations is lower, we observe a similar behavior as in Fig. 1. The type I error control is perfect for the Flipscores approach, even for small sample sizes, while the parametric tests show an anti-conservative behavior, the Wald test being better than the alternatives.

Figure 3 shows a power comparison. The true β is set equal to 0.5 while we test $H_0 : \beta = 0$, whose corresponding variable X is assumed continuous. The parameter γ is set equal to 0.5. The parametric tests are slightly more powerful - which is somehow expected from the anti-conservative behavior on the type I error control - but the loss of

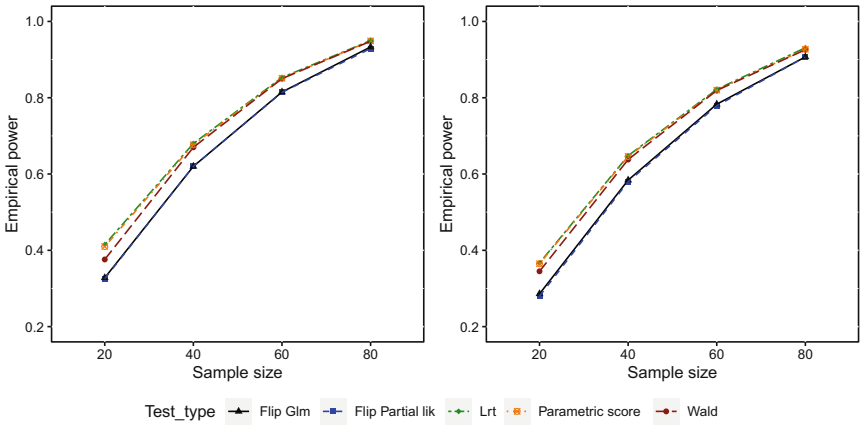


Fig. 3. Power comparison for continuous covariate. Average censoring equal to 25%.

power for the Flipscores approach is appreciably small, being almost equal for the two methods.

References

1. Arboretti, R., Fontana, R., Pesarin, F., Salmaso, L.: Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring. Stat. Methods Med. Res. **27**(12), 3739–3769 (2018)
2. Bender, R., Augustin, T., Blettner, M.: Generating survival times to simulate cox proportional hazards models. Stat. Med. **24**(11), 1713–1723 (2005)
3. David, CR.: Regression models and life tables (with discussion). J. Roy. Stat. Soc. Series B **34**, 187–220 (1972)
4. David, C.R.: Partial likelihood. Biometrika, **62**(2), 269–276 (1975)
5. De Santis, R., Goeman, J.J., Hemerik, J., Davenport, S., Finos, L.: Inference in generalized linear models with robustness to misspecified variances. arXiv preprint arXiv:2209.13918 (2022)
6. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. Roy. Stat. Soc. Series B-stat. Meth. **82**(3), 841–864 (2020)
7. Holford, T.R.: The analysis of rates and of survivorship using log-linear models. Biometrics **36**(2), 299–305 (1980)
8. Johansen, S.: An extension of cox's regression model. Int. Stat. Rev. **51**, 258–262 (1983)
9. Klein, J.P., et al.: Survival analysis: techniques for censored and truncated data. vol. 1230. Springer (2003). https://doi.org/10.1007/b97377
10. Laird, N., Olivier, D.: Covariance analysis of censored survival data using log-linear analysis techniques. J. Am. Stat. Assoc. **76**(374), 231–240 (1981)
11. Murphy, S.A., Van der Vaart, A.W.: On profile likelihood. J. Am. Stat. Assoc. **95**(450), 449–465 (2000)
12. Sun, Y., Sherman, M.: Some permutation tests for survival data. Biometrics, 87–97 (1996)

The Distribution of the Risk Function for Interval Estimation

Fulvio De Santis, Stefania Gubbiotti, and Francesco Mariani(✉)

Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
{fulvio.desantis,stefania.gubbiotti,f.mariani}@uniroma1.it

Abstract. In frequentist statistical decision theory, the average performance of an interval estimator is measured by its risk function, that is the expected value of the loss associated to the estimator. Risk functions depend on the unknown parameter of the model and might be summarized by the Bayes risk, their expected value with respect to a distribution assigned to the parameter. In this article we propose to study the entire distribution of the risk function induced by the prior and to complement the Bayes risk with additional summaries, such as the median and the mode, in order to define suitable criteria for sample size determination. We consider a standard loss, that is a monotone function of the interval length. We determine closed-form results for the Poisson-Gamma model and we illustrate an application to an oral health intervention study.

Keywords: Bayes risk · confidence intervals · decision theory · sample size determination

1 Methodology

Let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ denote a random sample and $f_n(\cdot|\theta)$ its probability mass or density function. Suppose we are interested in set estimation of the unknown parameter $\theta \in \Omega$ and, for simplicity, assume that $\Omega \subseteq \mathbb{R}$, i.e. that θ is scalar. Suppose also that any candidate set estimator of θ is an interval, defined as

$$C(\mathbf{X}_n) = [L(\mathbf{X}_n), U(\mathbf{X}_n)],$$

where, for any observed $\mathbf{x}_n, L(\mathbf{x}_n) \leq U(\mathbf{x}_n)$. Under the frequentist decision-theoretic framework, each interval C is evaluated in terms of its risk function, defined as

$$R_n(\theta, C) = \mathbb{E}_\theta[\mathbb{L}(\theta, C(\mathbf{X}_n))],$$

where $\mathbb{L}(\theta, C(\mathbf{X}_n))$ is the loss function of $C(\mathbf{X}_n)$ and $\mathbb{E}_\theta[\cdot]$ denotes the expectation wrt $f_n(\cdot|\theta)$. Typically, interval estimators are chosen within the class of $1-\gamma$ confidence sets, where $1-\gamma$ is the chosen confidence level (e.g., $1-\gamma = 0.95$). In this case the most common family of loss functions is given by the class of monotone functions defined as $\mathbb{L}(\theta, C) = \mathbb{S}[\mathcal{L}(\mathbf{X}_n)]$, where $\mathbb{S}[\cdot]$ is a non-decreasing function of $\mathcal{L}(\mathbf{X}_n) = U(\mathbf{X}_n) - L(\mathbf{X}_n)$, the random length of $C(\mathbf{X}_n)$. The risk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 543–548, 2025.
https://doi.org/10.1007/978-3-031-64431-3_90

function is the basic tool for evaluation of the performance of a single estimator, for the comparison of competitive sets and for the search of optimal solutions. However, it is a function of θ and this fact makes it difficult to compare alternative estimators since, typically, their risk functions intersect as θ varies in Ω. For this reason, risk functions are usually condensed into single numbers that can properly represent the overall average loss of an estimator.

One standard way to summarize the risk function of an interval estimator C into a single number is to treat the unknown parameter of the model as a random variable, Θ, with probability distribution $\pi(\cdot)$, and to consider the Bayes risk, that is the expected value of the random risk $Y_n = R_n(\Theta, C)$ with respect to $\pi(\cdot)$:

$$r_\pi(C) = \mathbb{E}_\pi[Y_n] = \mathbb{E}_\pi[R_n(\Theta, C)].$$

However, the Bayes risk is not necessarily a good summary of the random risk. The goal of this article is to show that inspection of the whole distribution of Y_n might be useful to assess whether $r_\pi(C)$ is an adequate summary or not and to suggest alternatives to it. We also want to show the usefulness of the proposed approach for sample size determination. To elaborate on these ideas, in the following section we derive the cumulative distribution function (cdf) $G(\cdot)$ and density function (pdf) $g(\cdot)$ of Y_n for the Poisson-Gamma model. We exemplify the methodology in the context of an oral health intervention study. In Sect. 3 we consider the sample size determination (SSD) problem and in Sect. 4 we sketch some concluding remarks.

2 Results for the Poisson-Gamma Model

Let $X_i|\theta \sim \text{Pois}(\theta)$, $i = 1, \ldots, n$ (i.i.d.), $\theta > 0$. The standard Wald $1 - \gamma$ confidence interval for θ is $C = \bar{X}_n \pm z_{1-\gamma/2}\sqrt{\bar{X}_n/n}$. For mathematical convenience, let $\mathbb{S}[\cdot]$ be the square function. Hence

$$\mathbb{L}(\theta, C) = \mathcal{L}(\mathbf{X}_n)^2 = a_n \bar{X}_n, \quad a_n = \frac{4}{n} z^2_{1-\gamma/2}.$$

If we consider Θ as a random variable, then $Y_n = R_n(\Theta, C) = a_n\Theta$. Let $\mathbb{F}_\pi(\cdot)$ and $\pi(\theta)$ be the cdf and pdf of Θ. It is easy to check that

$$G(y) = \mathbb{F}_\pi\left(\frac{y}{a_n}\right) \quad \text{and} \quad g(y) = \frac{1}{a_n}\,\pi\left(\frac{y}{a_n}\right), \quad y \geq 0.$$

It is also straightforward to verify that

$$\mathbb{E}_\pi(Y_n) = a_n\mathbb{E}_\pi(\Theta), \qquad \mathbb{M}\text{e}_\pi(Y_n) = a_n\mathbb{M}\text{e}_\pi(\Theta), \qquad \mathbb{M}\text{o}_\pi(Y_n) = a_n\mathbb{M}\text{o}_\pi(\Theta).$$

In the following we assume that $\Theta \sim \text{Ga}(\alpha, \beta)$, $\alpha, \beta > 0$. Then

$$g(y) = \frac{\beta^\alpha}{a_n^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\beta\frac{y}{a_n}}, \quad y > 0$$

and

$$\mathbb{E}_\pi(Y_n) = \frac{a_n\alpha}{\beta}, \qquad \mathbb{M}\mathrm{e}_\pi(Y_n) = a_n q_{0.5}, \qquad \mathbb{M}\mathrm{o}_\pi(Y_n) = \frac{a_n(\alpha-1)}{\beta},$$

where q_ε is the ε-level quantile of a $\mathrm{Ga}(\alpha,\beta)$ random variable.

Example: oral health intervention study

Now we consider the setup of a longitudinal oral health intervention study based on a sample of 7-year-old children born in 1989 in Flanders, Belgium (see [4,5]). The endpoint is the dmft-index, that is the number of primary teeth that are decayed, missing due to extraction or filled for caries: this index ranges between 0 (no caries) - 20 (all primary teeth with caries). The data are assumed to be a realization of independent Poisson random variables whose unknown parameter θ is the average number of caries experienced in the population. Based on historical data, in [5] a $\mathrm{Ga}(\alpha = 3, \beta = 1)$ is considered as a prior for Θ, which corresponds to the information of a single sample observation. This prior has expected value equal to 3 and assigns a probability as small as 0.001 to values larger than 10. Its standardized third moment, that is the usual measure of the skewness, is $\bar{\mu}_3(\Theta) = 2/\sqrt{\alpha} = 1.15$. Figure 1 shows the plot of $g(y)$ for $n = 10$ (left panel) and $n = 20$ (right panel). The vertical lines are drawn at $\mathbb{E}_\pi(Y_n)$ (solid line), $\mathbb{M}\mathrm{o}_\pi(Y_n)$ (dashed line) and $\mathbb{M}\mathrm{e}_\pi(Y_n)$ (dotted line); numerical values are reported in Table 1(a). Discrepancy between these values is determined by the right skewness of $g(y)$, $\bar{\mu}_3(Y_n)$, which coincides with the skewness of $\pi(\theta)$. As a consequence, among the three summaries, the expectation gives the largest synthetic measure of risk. Differences reduce as n increases.

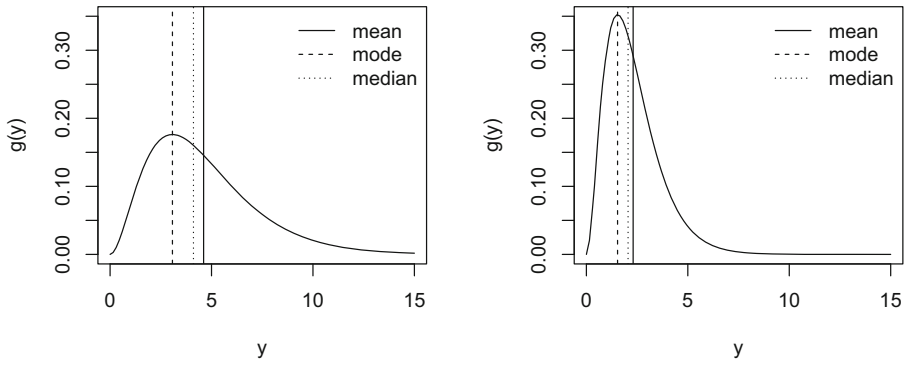


Fig. 1. Plots of $g(y)$ for $n = 10$ (left panel) and $n = 20$ (right panel), using the prior $\pi(\theta) = \mathrm{Ga}(\theta|\alpha = 3, \beta = 1)$. Vertical lines denote $\mathbb{E}_\pi(Y_n)$ (solid line), $\mathbb{M}\mathrm{o}_\pi(Y_n)$ (dashed line) and $\mathbb{M}\mathrm{e}_\pi(Y_n)$ (dotted line) respectively.

Following [5], after an interim look at the data (26 events observed on the first 10 children), we obtain an updated density $\text{Ga}(\alpha = 29, \beta = 11)$. Figure 2 shows the plot of $g(y)$ for $n = 10$ (left panel) and $n = 20$ (right panel). Note that, for both values of n, $g(y)$ is remarkably less skewed than in the previous cases, due to the lower skewness of the prior on Θ. This results in closer values of the three summaries (see Table 1(b)).

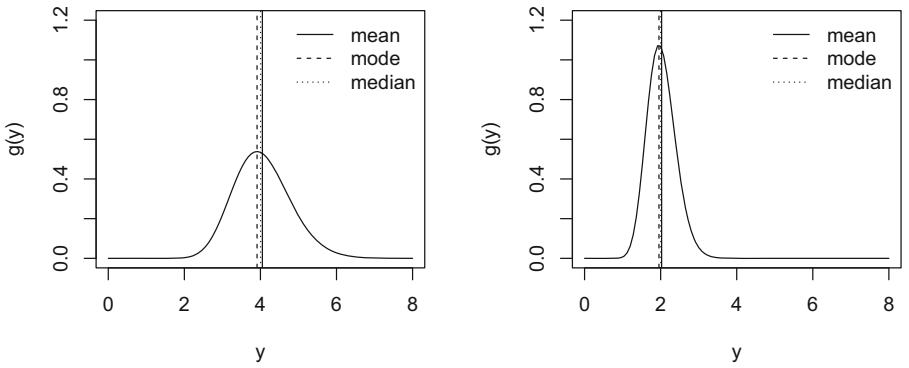


Fig. 2. Plots of $g(y)$ for $n = 10$ (left panel) and $n = 20$ (right panel), using the prior $\pi(\theta) = \text{Ga}(\theta|\alpha = 29, \beta = 11)$. Vertical lines denote $\mathbb{E}_\pi(Y_n)$ (solid line), $\mathbb{M}\text{o}_\pi(Y_n)$ (dashed line) and $\mathbb{M}\text{e}_\pi(Y_n)$ (dotted line) respectively.

Table 1. Summaries of $g(y)$ for several values of n using the priors (a) $\pi(\theta) = \text{Ga}(\theta|\alpha = 3, \beta = 1)$, (b) $\pi(\theta) = \text{Ga}(\theta|\alpha = 29, \beta = 11)$.

	n	$\mathbb{E}_\pi(Y_n)$	$\mathbb{M}\text{e}_\pi(Y_n)$	$\mathbb{M}\text{o}_\pi(Y_n)$	$\bar{\mu}_3(Y_n)$
(a)	10	4.61	4.11	3.07	1.15
	20	2.30	2.05	1.54	1.15
(b)	10	4.05	4.00	3.91	0.37
	20	2.03	2.00	1.96	0.37

3 Sample Size Determination

We can now use the distribution of Y_n to define SSD criteria. We assume that $\lim_{n\to\infty} Y_n = \lim_{n\to\infty} R_n(\theta, C) = 0$, $\forall\theta \in \Omega$. This implies that, as $n \to \infty$, $R_n(\Theta, C) \to 0$ pointwise. Note that consistency of Y_n implies that the three quantities $\mathbb{E}_\pi(Y_n)$, $\mathbb{M}\text{e}_\pi(Y_n)$ and $\mathbb{M}\text{o}_\pi(Y_n)$ tend to 0 as n tends to infinity. Hence, for a chosen threshold λ, we define the following optimality criteria for SSD:

$$\begin{aligned} n^\star_{ex} &= \min\left(n \in \mathbb{N} : \ \mathbb{E}_\pi(Y_n) \leq \lambda\right), \\ n^\star_{me} &= \min\left(n \in \mathbb{N} : \ \mathbb{M}\text{e}_\pi(Y_n) \leq \lambda\right), \\ n^\star_{mo} &= \min\left(n \in \mathbb{N} : \ \mathbb{M}\text{o}_\pi(Y_n) \leq \lambda\right). \end{aligned} \tag{1}$$

Example: oral health intervention study (continued)

Let us go back to the setup of the previous example and suppose we want to design a new study aimed at estimating the expected value of the dmft-index for the population of 7-year-old children. We want to determine the minimum sample size such that a chosen summary of the risk function (that is the expectation of the squared length of the 95% confidence interval) is sufficiently small. Figure 3 shows the plots of $\mathbb{E}_\pi(Y_n)$ (solid line), $\mathbb{M}\text{o}_\pi(Y_n)$ (dashed line) and $\mathbb{M}\text{e}_\pi(Y_n)$ (dotted line) as functions of n, under the two prior assumptions considered above. As expected, the distance between the three curves is more remarkable when the (a) $\text{Ga}(\theta|\alpha = 3, \beta = 1)$ is used (left panel). Table 2 reports the optimal sample sizes obtained according to criteria (1). As a consequence of the right skewness of the distribution of Y_n, for both priors, we have $n^\star_{ex} > n^\star_{me} > n^\star_{mo}$. Consistently with the behavior highlighted in Fig. 3 (right panel), the optimal sample sizes are almost coincident when the prior (b) $\text{Ga}(\theta|\alpha = 29, \beta = 11)$ is used.

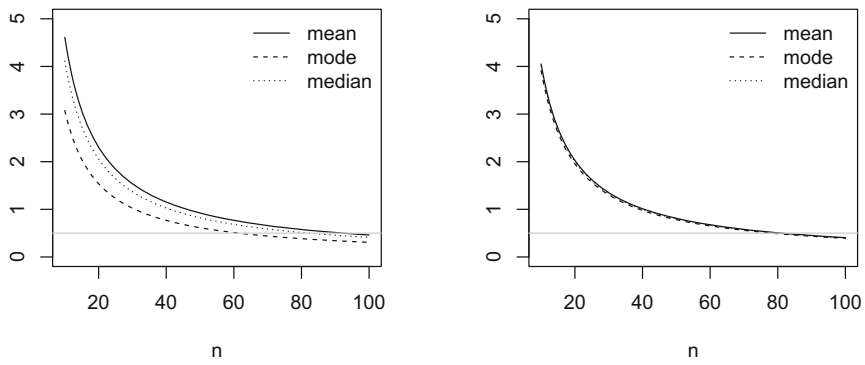


Fig. 3. Plots of $\mathbb{E}_\pi(Y_n)$ (solid line), $\mathbb{M}\text{o}_\pi(Y_n)$ (dashed line) and $\mathbb{M}\text{e}_\pi(Y_n)$ (dotted line) as functions of n, using the priors (left panel) $\pi(\theta) = \text{Ga}(\theta|\alpha = 3, \beta = 1)$, (right panel) $\pi(\theta) = \text{Ga}(\theta|\alpha = 29, \beta = 11)$.

Table 2. Optimal sample sizes $n^\star_{ex}$, $n^\star_{me}$ and $n^\star_{mo}$, using the priors (a) $\pi(\theta) = \text{Ga}(\theta|\alpha = 3, \beta = 1)$, (b) $\pi(\theta) = \text{Ga}(\theta|\alpha = 29, \beta = 11)$.

	$n^\star_{ex}$	$n^\star_{me}$	$n^\star_{mo}$
(a)	84	74	53
(b)	73	72	70

4 Concluding Remarks

In this article we propose to study the distribution of the risk function of $1-\gamma$ confidence intervals when the loss is a monotone function of their length. As an illustrative example, we consider the Wald interval for the parameter of the Poisson model and we derive explicit expressions for cdf and pdf of $R_n(\Theta, C)$. We show the crucial effect of the prior $\pi(\theta)$ on the shape of $g(y)$; in particular, since $\bar{\mu}_3(\Theta) = \bar{\mu}_3(Y_n)$, when the prior is informative enough and fairly symmetric with respect to its mode, $\mathbb{E}_\pi(Y_n)$, $\mathbb{M}\text{o}_\pi(Y_n)$ and $\mathbb{M}\text{e}_\pi(Y_n)$ tend to be numerically very close. This fact has also an impact in terms of sample size determination: a skewed prior implies different optimal sample sizes based on the three summaries of the risk, with $n^\star_{ex}$ quite larger than $n^\star_{me}$ and $n^\star_{mo}$; whereas a prior with a small value of $\bar{\mu}_3(\Theta)$ yields very similar optimal sample sizes.

The analysis proposed in the article is closely related to the one presented by [1] in the context of testing hypotheses. See also [2,3]. This approach can be easily extended to other models and inferential problems. When closed-form expressions for $g(y)$ and its summaries are not available, one can resort to straightforward Monte Carlo computations. In addition, for more accurate evaluations of the risk associated to a procedure, alternative summaries of $g(y)$ could be considered, such as credible intervals for $R_n(\Theta, C)$. These are topics for future research.

References

1. Rufibach, K., Burger, H., Abt, M.: Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. Pharm. Stat. **15**, 438–446 (2016)
2. De Santis, F., Gubbiotti, S.: On the limit distribution of the power function induced by a design prior. Stat. Papers **65**, 1927–1945 (2024). https://doi.org/10.1007/s00362-023-01462-9
3. De Santis, F., Gubbiotti, S., Mariani, F.: On Bayesian power analysis in reliability In: Chelli, F. M., Ciommi, M., Ingrassia, S., Mariani, F., Recchioni, M. C. (eds.) Statistical Learning, Sustainability and Impact Evaluation, Book of the Short Papers SIS, pp. 918-922. Pearson (2023)
4. Vanobbergen, J., L. Martens, E. Lesaffre, Declerck D.: Caries prevalence in Belgian children: A review. International Journal of Paediatric Dentistry, vol. 11, pp. 164–70 (2001)
5. Lesaffre, E., Lawson A. B.: Bayesian Biostatistics (2012). Hoboken, New Jersey: Wiley

Relational Hypervents Models for Co-authorship Networks: The Case of the Italian Academic Statisticians

Domenico De Stefano, Amin Gino Fabbrucci Barbagli(✉), Francesco Santelli, and Susanna Zaccarin

University of Trieste, Trieste, Italy
{ddestefano,fsantelli}@units.it, amingino.fabbruccibarbagli@phd.units.it, susanna.zaccarin@deams.units.it

Abstract. In the last year, there has been a growing emphasis on the significance of scientific collaboration. Several studies have analyzed co-authorship networks as a convenient tool to investigate collaboration among scientists across different disciplines and over time. Co-authorship networks among scientists can be represented by hypergraphs, where each scientific paper is the corresponding hyperevent. To account for this scenario and evaluate the evolution of the co-authorship network over time, Relational Hypervent Models (RHEM) have been recently proposed. Such models can handle time-varying data and events that occur in (weighted) hyperedges. They represent a suitable class of models for co-authorship data as they can model events composed of any authors, analyze polyadic settings, and consider sub-group persistence over time. This paper aims to apply such models to analyse the evolution of collaboration among the Italian academic statisticians observed over 10 years, from 2012 to 2022, including their external co-authors. Bibliographic information has been retrieved from the Scopus database. Different model specifications are proposed to evaluate the tendency to maintain collaboration among authors over time (familiarity effects) and/or the tendency to create new collaborations with authors who co-authored with common groups of authors (closure effects).

Keywords: Relational Hypervent Models · Scientific collaboration · Co-authorship · Hypergraph

1 Introduction

Recently, there has been a growing emphasis on the significance of scientific collaboration as it offers a valuable opportunity to leverage diverse expertise. It has been usually studied as a co-authorship process that allows researchers to share ideas, knowledge, and techniques, thus increasing scientific productivity.

The high availability of bibliometric data and its accessibility in platforms such as Scival, Scopus, Google Scholars (GS), and Web of Science (WOS) allow

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 549–554, 2025.
https://doi.org/10.1007/978-3-031-64431-3_91

for retrieving a large quantity of information related to productivity and production. Several works have proposed network analysis as a convenient tool to study co-authorship [7,9].

The evolution of co-authorship networks can be represented as hypergraphs over time. Hypergraphs are a generalization of graphs [2], consisting of a set of nodes representing the authors and a set of hyperedges representing the scientific papers, which are in turn a different subset of nodes. In particular, for any author is known how many scientific papers she/he has authored/coauthored across the years and how many coauthors she/he had on each paper. This information can be used to weigh the strength of collaboration ties [10] and its impact on the maintenance (persistence) of the collaboration over time.

To analyze co-authorship networks over time, [3,4] proposed the Relational Hyperevent Model (RHEM), a new class of statistical models capable of handling time-varying data and events that occur in (weighted) hyperedges. This approach is effective for analyzing multiple publications with a varying number of authors in different years. It estimates different effects related to group persistence or complex interactions (as closure or transitivity in a hypergraph setting).

In this contribution, following [3] we propose different specifications of RHEM providing empirical results on the evolution of co-authorship networks of Italian academic statisticians from 2012 to 2022.

2 Relational Hyperevent Models

The Relational Hyperevent Model (RHEM) is a recent family of statistical models useful to assess the propensity of actors to continue their interaction over time, also accounting for different author characteristics [3]. The model has shown its adaptability to polyadic settings and large networks. In particular, RHEMs can deal with fine-grained time-stamped events as in co-authorship networks considering papers published in different years [4]. It can assess the propensity to co-participate in joint events (papers) given previous past events and/or covariates [8].

More specifically, scientific publications can be seen as a two-mode network, having a set of nodes (papers) and a second set of nodes (authors) or alternatively as a hypergraph [2] (see Fig. 1).

Formally, a hypergraph $G = (V, H)$ is given by a set of nodes V and a set of hyperedges $H \subseteq P(V)$. Each hyperedge $h \in H$ is a subset of nodes $h \subseteq V$ of any size where $|h|$ represents the cardinality. A relational hyperevent $e \in E$, representing a publication event in a co-authorship network (i.e., a paper), is given by a tuple $e = (t_e, h_e, x_e)$, where t_e is the time of the event (i.e., the publication date), $h_e \subseteq V_{t_e}$ is the hyperedge on which the event occurs, representing the authors of the paper, x_e is the event type and/or event weight [4, p. 3].

The ***event rate*** (*hazard rate or intensity*) [3] on h at time t given the network with the past events is defined as:

$$\lambda(t; h; G[E; t]) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(t \leq T \leq t + \Delta t | h_e = h \wedge t \leq t_e < t + \Delta t\})}{\Delta t} \qquad (1)$$

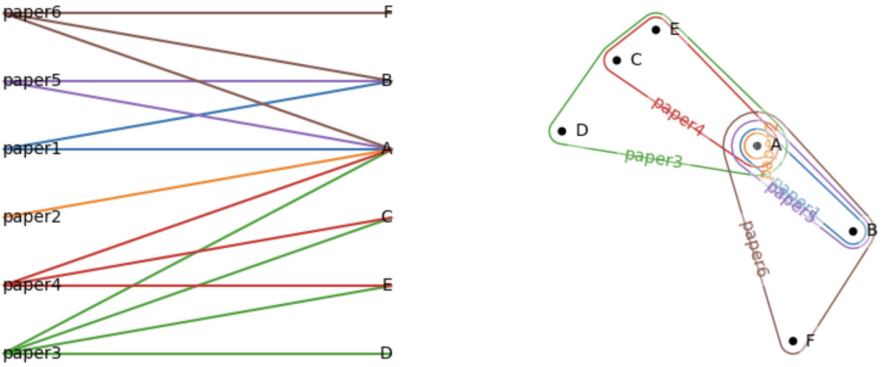


Fig. 1. Two-mode network of papers-authors in a given period (*Left*). Visualization as a hypergraph using the Python library HyperNetX (*Right*); each paper is considered a hyperevent.

Following the Cox proportional hazard model the event rate, λ can be decomposed into:

$$\lambda(t, h, \theta, G[E; t]) = \lambda_0(t) \cdot \lambda_{(1)}(t, h, \theta, G[E; t]) \tag{2}$$

$$\lambda_1(t, h, \theta, G[E; t]) = exp(\sum_{n=1}^{k} \theta_i \cdot s_i(t, h, G[E; t])) \tag{3}$$

with λ_0 is a baseline rate typically left unspecified which is constant for all hyperedges, λ_1 is the relative event rate, which is proportional to the probability that the event at time t occurs on the hyperedge h, $s_i(t, h, G[E; t])$ is a vector of statistics that indicates how the hyperedge h is embedded into the network of past events and θ is a vector of associated parameters describing which of these statistics increase or decrease the realtive event rate λ_1. Based on the observed event sequence E, the ***Partial likelihood*** is given by:

$$L(\theta) = \prod_{e \in E} \frac{\lambda_1(h_e; t_e, \theta; G[E; t])}{\sum_{h \in R_{te}} \lambda 1(h, t_e, \theta, G[E, t_e])} \tag{4}$$

Given the values of the statistics $s_i(t, h, G[E; t])$ for all elements of the risk sets R_{te} at the event times t_e, maximum likelihood estimates for Eq. (4) can be computed with standard statistical software [3].

It has been shown that this process can require a discrete amount of time given the size of the risk set, which grows exponentially with the number of nodes [5]. As an example, dealing with about 30,000 scholars and with a maximum number of authors per paper equal to 100 gives a risk set of $\binom{30000}{100}$ subsets.

The solution that has been proposed is to apply case-control sampling to ensure computational tractability. Then, the risk set R_{te} is further restricted to only hyperedges containing all participants lists of the same size as the

observed event, obtaining $R_{te} = h \subseteq V; |h| = |h_e|$, thus estimating the so-called conditional-size RHEM[4]. In our application, we focus on different RHEM specifications to analyze the tendency to maintain collaboration among authors over time (familiarity) and/or the tendency to create new collaborations with authors who collaborated with common groups of authors (closure). To this aim, we consider the following three simple effects (statistics).

Closure [5][3], that is, given two authors u and v who have previously worked with at least one common author, assesses the propensity of these two actors to interact directly [8].

$$closure(t, h', G[E,t]) = \sum_{u,v \in \binom{h}{2} \wedge w \neq u,v} \frac{min[deg(t,u,w;G[E,t]), deg(t,v,w;G[E,t])]}{\binom{|h|}{2}} \tag{5}$$

A positive closure suggests that there is a tendency to start work between authors who have not been authoring before (that is, actors u and v start to work together) and that dense overlapping hypergraphs tend to merge [4]. On the other hand, a negative closure suggests that the two authors, u and v, who share a common author, will not start working together, suggesting that dense, overlapping hypergraphs stay stable without merging [8].

The *subset repetition* [3] of order p takes into analysis an exact number of actors p that participated in the previous hyperevents, and it returns their propensity to participate again in a joint event in the future.

$$sub.rep^{(p)}(t, h, G[E,t]) = \sum_{h \in \binom{h}{p}} deg(t, h, G[E,t]) \frac{1}{\binom{|h|}{p}} \tag{6}$$

For example, the order $p = 1$ indicates the propensity of an individual author to continue her/his activity in the future (general attitude to publish), $p = 2$ suggests the tendency of a dyad of authors to participate again in a joint event, $p = 3$ of a triad [3] [4] and so on.

Repetition accounts for the tendency to repeat hyperevents involving the identical set of participants [3].

To account for the number of co-authors of a paper, we specify a conditional-size model [3]. As mentioned above, this strategy also allows for controlling the cardinality of the risk set. Alternatively, the latter aspect can be achieved by specifying the model only on repeated events that are those hyperedges that have experienced at least one event before t.

3 Data and RHEM Results

The above RHEM formulation is applied to co-authorship data on the Italian academic statisticians (IAS), as listed in the Italian Ministry of University roster in December 2022. Academic statisticians fall under the SECS-S sector and, more specifically, they are grouped into five (5) subsectors denoted

by the acronym S.S.D (Settore Scientifico Disciplinare). The resulting dataset comprised 836 scholars: SECS-S/01-Statistics (494), SECS-S/02-Statistics for Experimental and Technological Research (22), SECS-S/03-Economic Statistics (164), SECS-S/04-Demography (73), SECS-S/05-Social Statistics (83). Bibliometric information has been retrieved by querying the Scopus database from 2012 to 2022 using the authors' Scopus ID. Scopus ID is a unique identifier associated with authors provided in Scopus, thus avoiding misspellings and homonymy in retrieving the target authors and their production. The bibliographic database comprised more than 12,000 scientific publications, including additional information such as references, abstracts, and keywords, with an overall coverage rate among IAS close to 99%.

Results from different RHEM specifications are reported in Table 1.

Table 1. Results of different RHEM specifications

	Conditional size	Repeated events
repetition.avg.order.1	0.0072 (0.0003) ***	-0.0037 (0.0007) ***
repetition.avg.order.2	0.0782 (0.0026) ***	0.0602 (0.0071) ***
repetition.avg.order.3	0.0565 (0.0048) ***	0.0457 (0.0121) ***
repetition		0.0528 (0.0055) ***
closure	-0.0003 (0.0000) ***	-0.0015 (0.0004) ***
number.of.coauthors		-0.3509 (0.0232) ***
AIC	182227.1758	27659.0540
Num. events	12485	2477
Num. obs.	24006	4511

*** $p < 0.001$

The models show the IAS propensity to continue (or not) publishing in different settings, uncovering the tendencies of the target groups to collaborate over time (see Table 1). The positive values for all the repetitions of the model conditioned on size suggest the authors' propensity to co-participate again in a future joint event and to continue publishing in dyadic and triadic settings ($p = 2$, $p = 3$). The repetition.avg.order.1 shows the authors' tendency to continue individual activity, regardless the co-authors. On the other hand, the repeated events model shows that the tendency to produce new papers is negatively related to the number of co-authors with whom they previously worked. In such a model, we also observe a negative repetition of the order 1 effect (i.e., individual activity), partially compensated by the positive effect of repetition coefficient. On the other hand, the dyadic and triadic settings are positive and significant, showing the authors' tendency to appear jointly in future events.

The closure is significant and negative in both models, indicating that groups might stay stable and tend not to work with the co-author of the co-authors and that hyperevents do not merge but stay stable.

These findings suggest that smaller groups might be more effective in certain situations. Moreover, the presence of external co-authors is one of the main limitations of the present study. Furthermore, these results can be affected by the inner boundaries we set for our networks. In particular, since we collected bibliographic data on the target population of the IAS, the bibliographic records of co-authors external to this scientific community are not available.
This is a preliminary use of the Relation hypervent models, and in future works, we will propose a weighted version of RHEM that can be able to capture the presence of different types of external co-authors and their effect on the evolution over time of the collaboration among IAS. Since in co-authorship networks, weights can account for the strength of collaborative ties or the involvement of individual scientists in the joint scientific activity behind a paper's writing. Furthermore, we will explore the effect of other hyperevent weights, such as the impact factor or h-index of the journal where the papers have been published, the total amount of citations per year, or other metrics related to the published papers. The combinations of the above-mentioned weights could give new insights into the propensity to collaborate depending on the paper's characteristics.

References

1. Borgan, O., Goldstein, L. & Langholz, B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals Of Statistics*. pp. 1749-1778 (1995)
2. Bretto, A. Hypergraph theory. *An Introduction. Mathematical Engineering. Cham: Springer*. **1** (2013)
3. Lerner, J., Tranmer, M., Mowbray, J. & Hancean, M. REM beyond dyads: relational hyperevent models for multi-actor interaction networks. ArXiv Preprint ArXiv:1912.07403. (2019)
4. Lerner, J., Hâncean, M.: Micro-level network dynamics of scientific collaboration and impact: relational hyperevent models for the analysis of coauthor networks. Network Science. **11**, 5–35 (2023)
5. Lerner, J., Hâancean, M., Lomi, A.: Relational hyperevent models for the coevolution of coauthoring and citation networks. J. R. Stat. Soc. Ser. A: Stat. Soc. (2024). https://doi.org/10.1093/jrsssa/qnae068
6. Lerner, J., Lomi, A., Mowbray, J., Rollings, N., Tranmer, M.: Dynamic network analysis of contact diaries. Social Networks. **66**, 224–236 (2021)
7. De Stefano, D., Kronegger, L., Sciabolazza, V., Vitale, M. & Zaccarin, S. Social network tools for the evaluation of individual and group scientific performance. *Teaching, Research And Academic Careers*. pp. 165 (2022)
8. Lerner, J. & Lomi, A. A dynamic model for the mutual constitution of individuals and events. *Journal Of Complex Networks*. **10**, cnac004 (2022)
9. De Stefano, D., Fuccella, V., Vitale, M., Zaccarin, S.: Quality issues in co-authorship data of a national scientific community. Network Science. **11**, 98–112 (2023)
10. Newman, M. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*. **64**, 016132 (2001)

Misogynistic Content Detection Over Social Networks

Emiliano del Gobbo[1(✉)], Elisa Ignazzi[2], Barbara Cafarelli[1], and Lara Fontanella[3]

[1] Department of Economics, Management and Territory, University of Foggia, Via da Zara, 11, 71121 Foggia, FG, Italy
emiliano.delgobbo@unifg.it

[2] Department of Neuroscience, Imaging and Clinical Sciences, G. d'Annunzio University of Chieti-Pescara, Via dei Vestini, 33, 66100 Chieti, CH, Italy

[3] Department of Legal and Social Sciences, G. d'Annunzio University of Chieti-Pescara, Viale Pindaro, 42, 65127 Pescara, PE, Italy

Abstract. The identification of misogynistic content on social networks presents numerous challenges for Natural Language Processing Techniques, primarily due to the intricate nature of its detection, which necessitates a heightened level of contextual awareness. Our research employs a dataset derived from major platforms such as Twitter, Facebook, and Instagram, focusing on posts related to prominent female figures in Italy. The dataset, comprising 16,500 messages, has been meticulously annotated by three independent annotators to discern misogynistic content. Leveraging advanced natural language processing techniques, the study aims to develop an effective model for the automatic identification of misogynistic language within the dynamic context of social media. The findings contribute to the ongoing discourse on mitigating online gender-based harassment, providing valuable insights for the development of robust content moderation mechanisms.

Keywords: misogyny · misogyny detection · natural language processing · statistical learning

1 Introduction

Social media has become a melting pot of thoughts and perspectives due to increased communication. However, regrettably, misogynistic content has also made its presence known. Misogyny has been studied by several scholars, with two of the most commonly adopted definitions being "misogyny is a cultural attitude of hatred for females because they are female" [5] and "misogyny is the system that operates within a patriarchal social order to police and enforce women's subordination and to uphold male dominance" [7].

This article explores the difficulties of recognising and responding to misogynistic language on social media. The internet sphere is filled with gender-based

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 555–560, 2025.
https://doi.org/10.1007/978-3-031-64431-3_92

discrimination, ranging from covert biases to overtly insulting statements. Having the ultimate goal to build a more welcoming and inclusive online community, it is imperative that we recognise misogynistic content and address the issue as we navigate the digital landscape.

The rest of the paper is structured as follows. Section 2 provides an overview of the relevant studies and research conducted in the field. Section 3 offers a summary of the dataset utilised in the execution of this study. Section 4 outlines the objectives of our research, discusses the challenges encountered, and presents the approach used for classification using BERT [2], a cutting-edge methodology for Natural Language Processing tasks. Section 5 presents initial findings and provides insight into the future development of this study.

2 Background

Current efforts focus on developing tools capable of effectively identifying and counteracting the spread of misogyny and sexism in virtual interactions, in order to ensure a more secure and safe space for all users. Among the various techniques, the most commonly used are pre-trained deep-learning models and multimodal models, machine learning algorithms (such as SVM, Naïve Bayes, or Random Forest) and Convolutional Neural Networks models. Some of the most valuable contributions to the automatic identification of misogynistic comments are evaluation campaigns such as EVALITA (Evaluation of NLP and speech tools for Italian), that include a task for Automatic Misogyny Identification (AMI) [1], IberLEF (Iberian Languages Evaluation Forum) featuring the EXIST task, that focuses on the Identification of sEXism in Social neTworks [10], SemEval (International Workshop on Semantic Evaluation) that proposed the MAMI task, designed specifically for Multimedia Automatic Misogyny Identification [3], and FIRE (Forum for Information Retrieval Evaluation) that includes the task of Arabic Misogyny Identification (ArMI) [8]. An additional noteworthy contribution on the subject concerns the creation of datasets annotated by experts, based on some of the most popular and widely used social media platforms. Among these, Reddit [4] and Gab [6] are certainly of interest, along with Twitter, recently rebranded as X [9].

3 Dataset

The analysed corpus was constructed through the collection of data from three prominent social networks: X (formerly Twitter), Facebook, and Instagram. Specifically, posts containing keywords related to renowned female figures in Italy were compiled from Twitter, while comments on Facebook and Instagram posts concerning noteworthy Italian women, such as politicians, artists, influencers, and activists, were included. The resulting dataset comprises 16,500 messages, ensuring a balanced representation across diverse social networks. To ascertain content characteristics, the dataset underwent annotation by three distinct annotators, who categorised the messages based on their content as either

misogynistic, exclusively offensive, or devoid of both attributes. Each comment was annotated by one woman and one man, selected from a pool of ten master's students in Social Research, along with a PhD student. In addition, the classification process entailed identifying the precise section of the message that exhibited a misogynistic context and categorising the type of misogyny based on a taxonomy derived from previous annotation schemes [1,4,12].

4 Methodology

The objective of the study is to identify misogynistic content contained in messages shared on social media platforms. This task poses multiple challenges due to the fact that misogyny frequently lacks explicit keywords. However, the identification of misogynistic content relies on the contextual analysis of the sentence. Hence, the methodology should tackle the problem by taking into account the complete sentences. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a state-of-the-art Natural Language Processing (NLP) methodology that has significantly advanced the field of language understanding and representation learning. Developed by Google researchers in 2018, BERT has demonstrated exceptional performance in various NLP tasks, including question answering, sentiment analysis, and language translation [2]. The main breakthrough of BERT lies in its ability to provide bidirectional contextualised representations of words or sub-word units. BERT, unlike traditional language models, employs a bidirectional approach during training, taking into account both preceding and following words to understand the complete context of a word. The bidirectional approach of BERT allows it to effectively capture complex syntactic and semantic relationships in a given context, resulting in improved accuracy and nuanced comprehension of language. This aspect is particularly significant given the nature of the task that this research is addressing. BERT is constructed based on the Transformer architecture, which has demonstrated remarkable efficacy in capturing extensive interdependencies within sequences. The attention mechanism of the Transformer enables the model to selectively concentrate on pertinent segments of the input sequence while making predictions, thereby facilitating the modelling of complex linguistic structures. The initial training phase of BERT consists of training the model on extensive collections of data using unsupervised learning objectives, such as predicting words that have been masked within a sentence or predicting the subsequent sentence in a document. The pre-training phase provides the model with a comprehensive comprehension of the language to model. In addition, BERT introduces the notion of attention masks, which involve randomly masking a subset of words in a sentence during training. The model is then given the task of predicting these masked words. This promotes the acquisition of contextual associations between words, thereby improving the capacity to comprehend the complexities of language. The pre-training process is a computationally demanding procedure that necessitates a substantial corpus and powerful computing capabilities. Therefore, several pre-trained models are available, and for this

research, the model "dbmdz/bert-base-italian-xxl-case" has been chosen. This model is trained on a large Italian corpus of 81GB and is case sensitive. After completing pre-training, the process of fine-tuning is carried out on particular downstream tasks, such as sentiment analysis, named entity recognition, or classification. By employing transfer learning, BERT is able to utilise the knowledge acquired during pre-training to achieve exceptional performance in a range of natural language processing (NLP) tasks, without the need for large amounts of task-specific labelled data. Our analysis involved refining the model specifically for classification purposes, utilising the dataset annotated for misogynistic content described in Sect. 3. The analysis was conducted using Python and the Transformers library [11].

5 Results and Conclusions

The initial findings of the trained model demonstrate promising results. The annotated dataset has been split in two parts, for training and validation. Respectively the 90% and 10% of the full dataset. The two parts have been split balancing the contents coming from the different social networks. Figure 1 shows the Confusion Matrix for the BERT model over the test dataset after

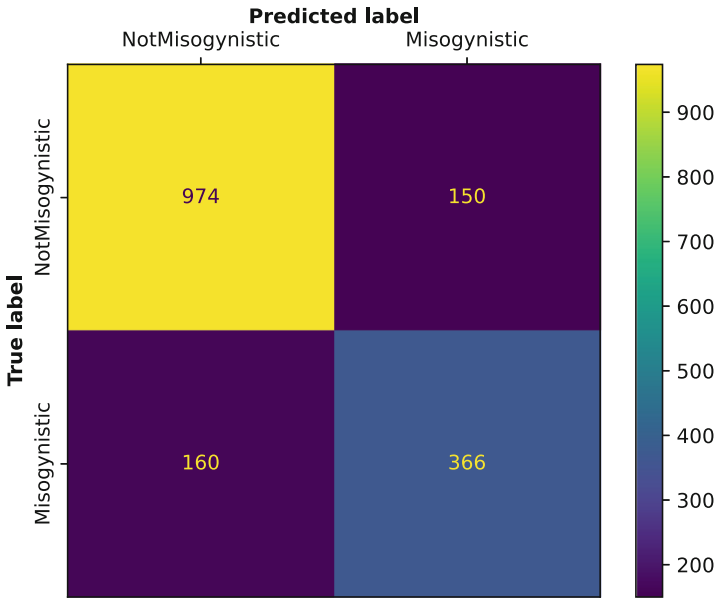


Fig. 1. Confusion Matrix of the preliminary model trained for the classification of misogynistic contents

fine-tuning the classification of misogynistic content using the training part of the dataset. The Accuracy score achieved by the model is 0.81. Table 1 displays summary measures of the goodness of fit of the trained model.

Table 1. Results of the preliminary model trained for the classification of misogynistic contents

Label	Precision	Recall	F1
Not Mysoginistic	0.87	0.86	0.86
Mysoginistic	0.70	0.71	0.70

As expected, the classifier's performance is higher for the non-misogynistic comments class, given the difficulty of identifying misogynistic content, especially where it is linked to neo-sexist discourses that use non-offensive language. Some examples of comments misclassified as not misogynous: *"[...] è la stessa disparità che esiste anche nel campo della disoccupazione: questa pesa negativamente (dal punto di vista dello status) enormemente più sull'uomo che sulla donna; un uomo, se è disoccupato, rischia l'emarginazione sociale (quella sessuale, è quasi scontata); per una donna è tutta un'altra cosa"*, *"[....] La donna secondo me deve sviluppare di più il suo lato razionale prendendo esempio dalle qualità maschili , disciplina nel lavoro ,forma di pensiero diretto e tanto altro [. . .]" "[...] Come ci sono quelle che si indignano dei commenti se l'uomo è brutto, ma ridono come oche se è bello o ha il macchinone.[. . .]"*

Based on these initial findings, future efforts will focus on utilising the unique characteristics of the annotated dataset. This includes utilising Part of Speech tagging to identify misogynistic content and determining the different types of misogyny present. Implementing a methodology that prioritises specific phrases for training and prediction might improve the accuracy of the results, offering more in-depth information about the content.

Acknowledgements. This research was undertaken as part of the ICOMIC (Identifying and Counteracting Online Misogyny in Cyberspace) Project funded by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021

References

1. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In: Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., and Meziane, F., (eds.)Natural Language Processing and Information Systems, pp. 57–64. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-91947-8_6

2. Devlin, J., Chang, M., 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, NAACL-HLT 2019,Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/N19-1423
3. Fersini, E., et al.: SemEval-2022 Task 5: multimedia automatic misogyny identification. In: Emerson, G., et al. (eds.) Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 533–549. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.semeval-1.74
4. Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., Margetts, H.: An expert annotated dataset for the detection of online misogyny. In: Merlo, P., Tiedemann, J., Tsarfaty, R., (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1336–1350, Online (2021). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.114
5. Johnson, M.P., Ferraro, K.J.: Research on domestic violence in the 1990s: making distinctions. J. Marriage Fam. **62**, 948–963 (2000). https://doi.org/10.1111/j.1741-3737.2000.00948.x
6. Kirk, H.R., Yin, W., Vidgen, B., Röttger, P.: SemEval-2023 task 10: explainable detection of online sexism. In: Ojha, A.K., Seza Doğruöz, A., Da San Martino, G., Madabushi, H.T., Kumar, R., Sartori, E., (eds.) Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pp. 2193–2210. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.semeval-1.305
7. Manne, K.: Down Girl: The Logic of Misogyny. Oxford Academic, New York (2017)
8. Mulki, H., Ghanem, B.: Working notes of the workshop Arabic misogyny identification (ArMI-2021). In: FIRE 2021: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 7–8. Association for Computing Machinery (2022). https://doi.org/10.1145/3503162.3503178
9. Pamungkas, E.W., Basile, V., Patti, V.: Misogyny detection in twitter: a multilingual and cross-domain study. Inform. Process. Manag. **57**(6), 102360 (2020). https://doi.org/10.1016/j.ipm.2020.102360
10. Rodriguez-Sanchez, F.J., et al.: Overview of EXIST 2021: sEXism Identification in Social neTworks. Proces. Leng. Nat. **67**, 195–207 (2021)
11. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Liu, Q., Schlangen, D., (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6
12. Zeinert, P., Inie, N., Derczynski, L.: Annotating online misogyny. In: Zong, C., Xia, F., Li, W., Navigli, R., (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3181–3197. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.247

Evaluation of Adjustment for Unobserved Confounders in Occupational Epidemiology Studies

Silvia D'Elia[1(✉)], Stefania Massari[2], Marco Alfò[1], Claudio Gariazzo[2], Alessandro Gialluisi[3], Alessandro Marinaccio[2], Licia Iacoviello[3], and Simona Costanzo[3]

[1] Department of Statistical Science, Sapienza University of Rome, Rome, Italy
silvia.delia@uniroma1.it
[2] Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, Italian National Institute for Insurance against Accidents at Work (INAIL),Rome, Italy
[3] Department of Epidemiology and Prevention, IRCCS Neuromed, Pozzilli, IS, Italy

Abstract. The issue of bias due to unmeasured confounding when estimating association between a given outcome and an exposure has been (and it is still) of great interest in occupational epidemiology. The main concern is that personal risk factors are lacking on administrative data, such as smoking. Different approaches, that require different degrees of available information, have been developed to evaluate the impact of the bias due to the failure in controlling for unmeasured confounders.Methodologies considered in the analysis not only include models for complete individual data, but also the indirect control through aggregate data, adding Monte Carlo simulation for the uncertainty of the estimates as well as techniques that do not require any information.These techniques are applied to a population cohort and results are compared to evaluate how the estimates vary across the different scenarios that have been proposed to adjust for confounding effect of smoking habit.

Keywords: occupational epidemiology · unobserved confounders · indirect control

1 Introduction

Confounding occurs when a variable is a risk factor for the outcome and it is also correlated with the exposure under investigation. Occupational epidemiology studies analyse occupational exposure-disease associations and commonly evaluate the excess risk of a particular disease in a specific occupational exposure class. The excess is assessed with respect to the general population, or to a population that is assumed not to be exposed or that can be supposed to have the same exposure of the general population.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 561–566, 2025.
https://doi.org/10.1007/978-3-031-64431-3_93

Occupational cohorts are usually assembled by record linkage of data from archives or administrative sources. Therefore, data rarely contains individual clinical and demographic information. Demographic characteristics are limited to the detection of features such as gender, age, residence, educational qualification or job role. Consequently the models are adjusted using the few variables available. It is uncommon for occupational cohort mortality studies to include smoking history information that often attracts considerable interest as alternative explanations or contributors to the excess morbidity or mortality associated with occupational exposure. Consequently, smoking is often an unmeasured potential confounder in the analyses.

The absence of information complicates interpretations since tobacco plays a major role in the etiology of many chronic diseases including cancer of the lung, esophagus, larynx, mouth and other sites as well as respiratory and cardiovascular disease. Therefore differences in smoking prevalence between specific occupational sectors may result in inaccurate assessment of the relationship between disease and the occupational exposure of interest [1].

Failure to consider individual features therefore leads to the hypothesis that all (unobserved) variables share the same distribution for each level of exposure. If this hypothesis does not hold, it is necessary to evaluate the impact of these factors in the calculation of the association. When unobserved confounders come into play in the relationship of interest, it is highly probable that the observed association between exposure and outcome is biased. However, even without direct information on their occurrence or magnitude in the population under study, the possible impact of confounders can be estimated [2].

The aim of the study is to present a review of techniques to adjust for confounders measures of association between outcome and exposure. Different methods with different degrees of information are evaluated, starting from models for complete individual data up to methods that may be applied in the absence of information regarding the confounder.

Using data from a population cohort where information on disease outcomes, occupational exposure and confounders is available, and applying techniques to control for unobserved confounders, it is possible to evaluate the impact of bias introduced by the failure to adjust for confounding, and compare the results obtained with methods characterized by different degree of information.

2 Material

The analysis is performed using the prospective cohort of the Moli-sani project [3]. Moli-sani collects data regarding an adult population cohort randomly drawn from the resident population in the Molise region, recruited in the 2005–2010 period ($N = 24325$; age $\geq$ 35 years; 52% women). Data was acquired from questionnaires relating to lifestyles, socio-economic level, biometric measures and clinical as well as laboratory data.

The cohort was merged via record linkage with the INPS archive containing data on work history ($N = 13971$). INPS files include individuals who worked in

private companies having at least one employee from 1974 onwards. The occupational exposure (E) is defined as the economic sector where the worker has been employed in most of their working life. The occupational sectors involved in the analysis are agriculture, buildings and construction industry, mechanical industry and the private service sector. Employees in the service sector are considered unexposed as their risks are similar to those of the general population.

The outcome (D) are the hospitalization events (coded by the ICD-9 classification), identified through the hospital discharge records (SDO) of the Molise region. Those lost to follow-up or subjects for whom demographic information was missing or not valid are excluded, with the aim to identify subjects hospitalized malignant tumor. The following cancer categories are considered: lymphatic and hematopoietic tissue, genitourinary organs, digestive system and respiratory system and intrathoracic organs. The confounding factor *smoking* (S) is a categorical variable with three levels: non-smoker, former smoker and current smoker.

3 Methods

3.1 Direct Control Using Individual Information

The exposure-outcome association is often estimated by the relative risk defined as the ratio between rates. The numerator of a rate can be considered as a realization of a Poisson variable. Consequently the relation between the rate and a variable (e.g. exposure E) can be investigated by a Poisson regression model.

$$\log(\widehat{\lambda}_E) = \hat{\beta}_0 + \hat{\beta}_1 E \tag{1}$$

$$\widehat{RR}_{ED}^{unadj} = \frac{\widehat{\lambda}_E}{\widehat{\lambda}_{\bar{E}}} = \exp(\hat{\beta}_1) \tag{2}$$

Let $\widehat{RR}_{ED}^{unadj}$ denote the estimated relative risk of the exposure E on the outcome D. In the presence of confounding it is convenient to consider a common estimate the effect of exposure across the strata of the confounder and to estimate the corresponding confidence intervals. Such a task may be performed by the Poisson regression model. Considering that the cohort allows to adjust for the confounder of interest (smoking S) with the individual information, it is possible to obtain the relative risk of the exposure E on the outcome D adjusted for the factor S. The confounder, in this case, is a categorical variable with three levels S_1, S_2, S_3. In the model below S_1 is set as a reference level.

$$\log(\widehat{\lambda}) = \hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 S_2 + \hat{\beta}_3 S_3 \tag{3}$$

The relative rate of the exposure is obtained by exponentiating the estimate of the exposure coefficient $\hat{\beta}_1$, which, according to eq. (3), results in an estimate of the relative risk adjusted for confounding $\widehat{RR}_{ED}^{adj}$.

3.2 Indirect Control Using the Bias Factor

The lack of appropriate individual data does not allow the calculation of the model in equation (3). It is possible to apply the model in eq. (1) in order to obtain a risk estimate and an associated confidence interval as a measure of uncertainty. However, this does not consider the bias that may arise from unobserved confounding. In the absence of smoking data, methods for external confounding adjustment have been described, that may produce a rough estimate of the potential confounding by smoking, see [4–6].

When it is not possible to access individual information, estimates can be adjusted using a bias factor; such procedure is based on the fact that the real magnitude of the association may be obtained calculating the ratio of the unadjusted one and the bias factor.

$$\widehat{RR}_{ED}^{B} = \frac{\widehat{RR}_{ED}^{unadj}}{B} \tag{4}$$

The bias factor may be calculated considering the distribution of the confounder (S) in the different exposure groups and the association between confounder and outcome given by the relative risk, based on the partitioning of the incidence rate proposed by Axelson [4].

$$B = \frac{P_{ES_1} + P_{ES_2} \cdot \widehat{RR}_{S_2D} + P_{ES_3} \cdot \widehat{RR}_{S_3D}}{P_{\bar{E}S_1} + P_{\bar{E}S_2} \cdot \widehat{RR}_{S_2D} + P_{\bar{E}S_3} \cdot \widehat{RR}_{S_3D}} \tag{5}$$

The bias factor is a ratio between a weighted average of the relative risks (outcome-confounder) for the prevalence of the various levels of the confounder in the exposed individuals, compared with the same weighted average for the unexposed ones.

The technique provides a quite precise estimate of the association between exposure and outcome accounting for the confounding factor, but not a measure of uncertainty to be associated with the estimate. A measure of uncertainty of this estimate may be provided via Monte Carlo simulation. Monte Carlo simulation also provides an overview of what happens when the values of the distribution of the confounder vary within the exposure groups. The simulation is based on the normal approximation of the logit of a proportion, and on the delta method for deriving the asymptotic variance of the estimator. $M = 1000$ generated values for the proportions of the confounder's levels in the exposed and unexposed groups, and M generated relative risks for the disease and the confounder, and for the disease and the exposure, are used to compute M bias factors using equation (5) above. Relative risks are generated form a normal distribution with a mean equal to the log of the original estimate, and a standard deviation equal to the standard error of that estimate. Each of these factors are used to adjust a sampled unadjusted relative risk $\widehat{RR}_{ED}^{unadj}$.

3.3 Control in the Absence of Information

Another method to assess how robust associations are to potential unmeasured or uncontrolled confounding is a measure called E-value [7]. The method does not require any individual information about the confounding factor. Without imposing any assumptions on the unmeasured confounder or confounders, a bounding factor and an inequality are derived.

$$RR_{ES} = RR_{SD} \geq \widehat{RR}_{ED}^{unadj} + \sqrt{\widehat{RR}_{ED}^{unadj}(\widehat{RR}_{ED}^{unadj} - 1)} \tag{6}$$

The E-value is defined as the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the exposure and the outcome to fully explain away a specific exposition-outcome association, conditional on the measured covariates.

4 Results

In Table 1 it is possible to verify that the distribution of the confounder *smoking* is not equal among the occupational sectors. Prevalence of current smokers is highest in construction (30%), service (28.4%) mechanical (27.7%). Lowest rate is observed among workers in agriculture (12.1%).

Table 1. Distribution of the confounder *smoking* in occupational sectors.

Sector	Non-smoker	Former	Current
Agriculture	70.2%	17.7%	12.1%
Mechanical	33.8%	38.4%	27.7%
Construction	28.2%	42%	29.8%
Service	45.5%	26.1%	28.4%

The results in 2 show that in sectors where the prevalence of smokers is lower, the adjusted relative risk increases significantly in correspondence with some categories of cancer (digestive system $\widehat{RR}_{adj} = 1.74\ CI95\% = [1.28; 2.35]$). The adjusted risk decreases in sectors where the prevalence of smokers is high (Table 2).

The bias-adjusted risk tends to get close to the adjusted estimates, denoting a good approximation while preserving information on the confounder.

The results obtained using the E-value show that the minimum association does not take on values that could be defined as unlikely, especially with regard to the confounding - outcome association.

Table 2. Relative risk unadjusted and adjusted using methods in pars- 2.1-2.3

Sector		$\widehat{RR}_{unadj}$	$\widehat{RR}_{adj}$	$\widehat{RR}^{B}_{ED}$	E
Agriculture	Lymphatic and hematopoietic tissue	1.12	1.13	1.11	1.49
	Genitourinary organs	0.91	1.09	1.09	1.41
	Digestive system	1.66	1.89	1.75	2.70
	Respiratory system and intrathoracic organs	1.16	1.85	1.80	1.59
Mechanical	Lymphatic and hematopoietic tissue	1.02	0.95	0.98	1.14
	Genitourinary organs	1.95	1.77	1.79	3.32
	Digestive system	1.45	1.40	1.37	2.27
	Respiratory system and intrathoracic organs	1.72	1.65	1.60	2.82
Construction	Lymphatic and hematopoietic tissue	0.92	0.82	0.88	1.39
	Genitourinary organs	1.80	1.58	1.59	3.00
	Digestive system	1.74	1.58	1.61	2.87
	Respiratory system and intrathoracic organs	2.70	2.47	2.37	4.84

5 Discussion

Smoking is widely recognised as one of the most important risk factor for different health outcomes. It is not possible to ignore it in epidemiologic analysis, given that smoking prevalence may significantly vary among the exposure levels [8].

It is always recommended to apply a technique that allows to use as much information as possible. If individual data is lacking, aggregate data from other sources is still valuable. It is important to evaluate the accuracy and reliability of the source from which the data is obtained.

A further assessment is to include the interaction between exposure and confounding in the models for complete data, when available. Future developments may refine smoking confounder definition and assess more confounding factors.

References

1. Blair, A., et al.: Control of smoking in occupational epidemiologic studies: methods and needs. Am. J. Ind. Med. **13**, 3-4 (1988)
2. Blair, A., et al.: Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. Am. J. Ind. Med. **50**, 199-207 (2007)
3. MOLI-SANI Study. www.moli-sani.org
4. Axelson, O.: Aspects on confounding in occupational health epidemiology. Scand. J. Work Environ. Health **4**, 98–102 (1978)
5. Axelson, O., Steenland, K.: Indirect methods of assessing the effects of tobacco use in occupational studies. Am. J. Ind. Med. **13**, 105–118 (1988)
6. Steenland, K., Greenland, S.,: Monte carlo sensitivity analysis and bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. Am. J. Epidemiol. **160**, 384–392 (2004)
7. Ding, P., VanderWeele, T.J.: Sensitivity analysis without assumptions. Epidemiology **27**, 368–377 (2016)
8. McCurdy, S.A., et al.: Smoking and occupation from the European community respiratory health survey. Occup. Environ. Med. **60**, 643–648 (2003)

Digitalization, Work-Related Risk Factors and Well-Being: Importance and Interactions from Tree-Based Methods

Fabio Demaria(✉) and Maddalena Cavicchioli

Department of Economics "Marco Biagi", University of Modena and Reggio Emilia, Viale Berengario 51, 41121 Modena, Italy
fabio.demaria@unimore.it

Abstract. In today's rapidly evolving workplace, the imperative to study workers' well-being has never been more apparent, considering both job-related risk factors and challenges posed by the digital transformation. Utilizing a large dataset from the recent European Working Conditions Telephone Survey (EWCTS) 2021, we explore the impact of physical and psychosocial risk factors on well-being and their interaction with the attributes of new forms of work brought by digitalization. Our methodological framework employs a suite of decision tree-based methods and sophisticated visual tools, designed to extract insights through a non-parametric approach. Findings reveal that increasing physical and emotional exhaustion and the need for recognition are predominant factors affecting employee well-being. Additionally, we find strong links between psychosocial risks, telework, and job locations.

Keywords: Employee well-being · Digital transformation · Random forest · Network graph · Variable interaction

1 Introduction

The ongoing digital transformation and its implications for organizational dynamics underscore the importance of human capital and its well-being [1]. International corporations are increasingly prioritizing human resource policies that foster employee well-being as a strategy to enhance engagement and, consequently, organizational performance [2]. The concept of well-being, introduced in the 1960 s in positive psychology, has gained importance both in social science research and public policy, as reflected by Sustainable Development Goal (SDG) 3, which focuses on global health and well-being [3,4]. The shift towards digital work, accelerated by the Covid-19 pandemic, has highlighted the importance of telework and digitalization in reorganizing social and work relations [5]. This study uses a decision tree-based framework to explore key factors affecting employee well-being and their association with work digitalization, drawing on data from the European Working Conditions Telephone Survey (EWCTS) 2021 conducted by Eurofound. This survey provides insights from around 70,000

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 567–573, 2025.
https://doi.org/10.1007/978-3-031-64431-3_94

workers across 36 European countries, reflecting on job quality amidst significant workplace transformations. The paper is structured as follows: Sect. 2 briefly outlines the theoretical background of the study; Sect. 3 describes the data and methodology; Sect. 4 discusses the results; and Sect. 5 concludes the study.

2 Theoretical Background

Well-being in the workplace encompasses employees' happiness, physical, and mental health [6]. Warr [7] defines job-related well-being as individuals' satisfaction with various job aspects, including pay, relationships with colleagues, interactions with supervisors, and working conditions. In psychology literature, well-being is primarily studied and measured as a psychological state. It is often referred to as subjective well-being (SWB), related to life satisfaction and personal condition assessments, and reflects the 'hedonic' orientation [8]. However, components of subjective well-being are not separate from other well-being components, such as self-realization and personal growth, elements related to the 'eudaimonic' orientation [9]. Thus, the two well-being orientations overlap and are interconnected [10]. In today's world, marked by the digital transformation of work practices, employee well-being is intrinsically influenced by the extent of digitalization within their occupational activities. Digital transformation in the workplace affects employee well-being, with technology potentially increasing flexibility and job enrichment but also possibly leading to higher job demands and work-life balance challenges [11].

3 Data and Methods

The sample study comprises 30,207 workers from the EU12, the original 12 European Union Member States, which exhibit similar political and socio-economic characteristics. We examine a total of 56 variables associated with eight groups identified in the EWCS-related literature [12,13]. *Job Responsibility* evaluates employees' involvement in organizational improvements. *Threats and Abuse in the Workplace* addresses experiences like verbal abuse. *Job Reward* measures perceptions of compensation fairness. *Physical* examines work environment factors, such as noise levels. *Psychosocial* assesses the balance between work and personal life. *Engagement* looks at motivational aspects, like energy at work. *Digital Work* explores the frequency of digital tool usage. *Demographics* gathers essential workforce data. The response variable is employee well-being, constructed as a dummy variable derived from scores on the *WHO-5 Well-Being Index* [14]. This index quantifies five dimensions of well-being using a 6-point Likert scale, from 0 (none of the time) to 5 (all of the time). The five items are: (1) 'I have felt cheerful and in good spirits', (2) 'I have felt calm and relaxed', (3) 'I have felt active and vigorous', (4) 'I woke up feeling fresh and rested' and (5) 'My daily life has been filled with things that interest me'. The composite score, ranging from 0 (no well-being) to 25 (optimal well-being) is then transformed into a percentage. Finally, a dummy variable is created using a 0.52 threshold [15], where

scores above 0.52 are coded as 1 (denoting higher well-being), and scores below as 0 (indicating lower well-being).

3.1 Methodology

Random Forest (RF) is an ensemble learning method for classification and regression introduced by Breiman [16]. It integrates Breiman's bagging approach and the random selection of features, a concept introduced by Ho [17]. The RF algorithm is adept at handling both continuous and categorical variables, identifying relevant variables, and interpreting complex interactions. In our study, we utilize both the supervised and unsupervised versions of RF. The supervised algorithm is employed to extract the most relevant variables, while the unsupervised version is used to investigate the patterns of relationships between the most significant variables and those related to digital work.

Variable Importance In this study, we leverage the RF algorithm for feature selection, utilizing the minimal depth concept as an innovative alternative to conventional methods like mean decrease impurity and permutation importance [16,18]. The minimal depth approach, which assesses the proximity of a variable to the tree's root, serves as an indicator of predictive strength—the lesser the minimal depth, the stronger the variable's predictive power [19]. This method is articulated through the concept of maximal subtree, calculated based on the tree's topology. Specifically, for a variable v, the first-order maximal subtree is the largest subtree in which v is utilized for the initial split at the root. The depth d of a node, denoting its distance to the root, falls within the set $\{0, 1, \ldots, D(T)\}$, where $D(T)$ signifies the tree's maximum depth. The minimal depth D_v for v is defined as the distance from the root to the nearest maximal subtree's root, where v first conducts a split, reflecting the predictive strength of v. The average of the minimal depth distribution acts as a threshold to discern relevant features, aligning with Ishwaran's methodology [19].

Variable Interaction The application of minimal depth extends to second-order maximal subtrees to examine variable interactions (Lu, 2021). In this context, a second-order maximal subtree $T_{v,w}$ within a maximal v-subtree T_v signifies an interaction: the proximity of w to v's root in $T_{v,w}$ indicates the strength of their relationship. The minimal depth q for $T_{v,w}$ is normalized by $D_{v,w}$, the depth of the relevant subtree, yielding a score between $[0, 1]$. This score, averaged over the entire forest, inversely correlates with the interaction strength. Employing maximal subtrees and minimal depths, a $p \times p$ interaction matrix $\mathbf{\Phi}$ is formulated, where p denotes the number of variables.

Network Graph To illustrate the interaction matrix from minimal depth analysis, we use a network plot, showcasing feature importance and interactions.

Larger, darker (red) nodes indicate more important features. Blue edges represent interactions between feature pairs, with thicker, more intense lines indicating stronger interactions. Features are arranged clockwise based on the DendSer algorithm, a seriation technique combining hierarchical clustering and sorting to position similar variables closer together, enhancing visualization clarity [20]. The seriation process relies on weights w_{ij} between object pairs i and j to reflect their relative importance, with lower weights signifying higher significance.

4 Results

The supervised RF classification model was employed, using the dichotomous well-being variable as the response and the eight variable groups as input. The model's performance indicates a classification error rate of 23.20%. Firstly, variable importance is evaluated through RF algorithm computing the minimal depth as described in Sect. 3 and results are reported in Table 1. Interestingly, we observe that employees' well-being is predominantly influenced by individual psychosocial factors such as work-life balance, work pace, and effort, which can lead to both emotional and physical exhaustion. These observations align with research on the effort-reward imbalance [21,22], underscoring the importance of balancing work demands and rewards to maintain mental health. Additionally, we note that the sole physical risk factor is closely linked to the emotional realm rather than to specific unhealthy work environments, confirming the crucial role of psychosocial factors. Secondly, we investigate well-being risk factors and their interactions with work digitalization, including tools, job locations, and the degree of telework. The 'telework' variable was adjusted to reflect our research objectives, with 1 indicating no telework and 6 representing full-time telework. The network plot in Fig. 1 shows that telework's degree is vital for psychological well-being, closely associated with feelings of emotional and physical

Table 1. Minimal depth rank.

Variable name	Code	Question description	Depth	Category
work_welldone	Q61H	[Please tell me how often the following applies to your work situation?] Your job gives you the feeling of work well done.	1.974	Psychosocial
eng_energy	Q90A	At my work I feel full of energy.	2.194	Engagement
work_life_balance	Q44	In general, how do your working hours fit in with your family or social commitments outside work?	2.252	Psychosocial
recognition	Q89D	I receive the recognition I deserve for my work.	2.372	Job reward
eng_enthusiastic	Q90B	I am enthusiastic about my job.	2.502	Engagement
wlb_tired	Q45B	[Since you started your main paid job, how often have you...?] Felt too tired after work to do some of the household jobs which need to be done.	2.512	Psychosocial
exhaust_emot	Q90G	[The following statements are about how you feel about your job. For each statement, please tell me how often you feel this way...] I feel emotionally exhausted by my work.	2.796	Psychosocial
exhaust_phys	Q90A	[The following statements are about how you feel about your job. For each statement, please tell me how often you feel this way...] I feel physically exhausted at the end of the working day.	3.342	Psychosocial
emot_disturb	Q30H	Being in situations that are emotionally disturbing for you.	3.418	Physical

exhaustion. Work-life balance also significantly influences teleworking choices, which in turn, is closely linked to well-being. While device usage is widespread, it is not a direct mental health stressor. The choice of job location, influenced by the shift toward remote work, also affects well-being. These insights are consistent with recent studies on telework and mental health [23], as well as the impact of work-life balance on well-being [13,24,25]. New work modalities, particularly those introduced in response to the pandemic, have heightened employees' concerns about working outside regular hours, thereby increasing fatigue and exhaustion.

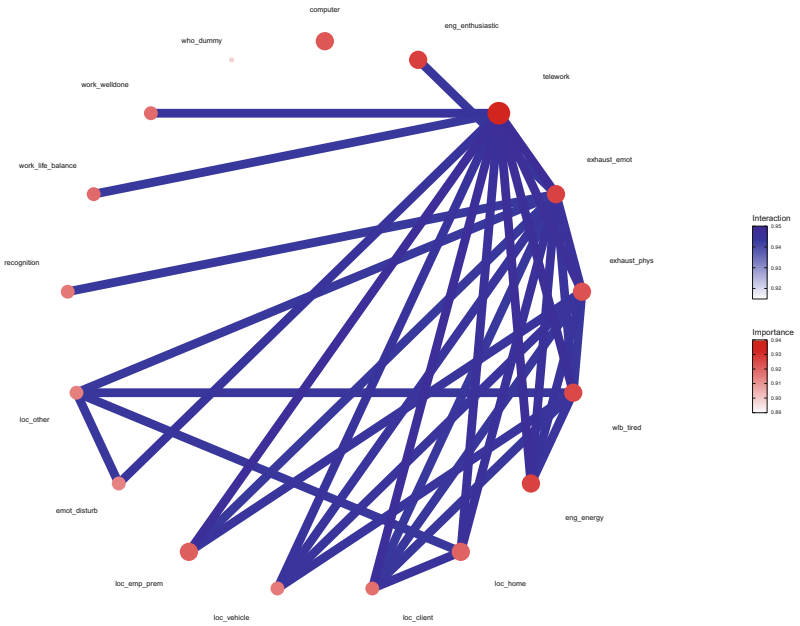


Fig. 1. Network graph.

5 Conclusions

In this work, we investigate the impact of work-related factors on employee well-being and their interactions with the digitalization of work, using recent tree-based methods and data from the EWCTS 2021 conducted by Eurofound. Specifically, we employ a set of both supervised and unsupervised decision tree-based methods to i) identify the most predictive features for employee well-being; ii) explore the relationships between the most relevant determinants of well-being and digital work. In this context, we identify psychosocial determinants as predominant, characterized by an increasing sense of exhaustion (physical and

emotional) that needs to be counterbalanced by recognition of effort and higher engagement. Furthermore, we address the analysis of various sets of occupational risks arising from the COVID-19 pandemic and the transformation of work due to the digitalization of workplaces. Notably, we find strong interconnections between major psychosocial risk factors and the degree of telework and job locations. Future research will investigate variations in network plots based on employee well-being levels and analyze geographical heterogeneity, to identify geographical patterns in workplace dynamics and well-being.

References

1. Chalutz Ben-Gal, H.: An ROI-based review of HR analytics: practical implementation tools. Pers. Rev. **48**(6), 1429–1448 (2019)
2. Gaston-Breton, C., Lemoine, J.E., Voyer, B.G., Kastanakis, M.N.: Pleasure, meaning or spirituality: cross-cultural differences in orientations to happiness across 12 countries. J. Bus. Res. **134**, 1–12 (2021)
3. Seligman, M., Csikszentmihalyi, M.: Positive psychology: an introduction. Am. Psychol. **55**(1), 5–14 (2000)
4. United Nations General Assembly. (2015). Transforming our world: The 2030. Agenda for sustainable development
5. Leonardi, P.M., Treem, J.W.: Behavioral visibility: a new paradigm for organization studies in the age of digitization, digitalization, and datafication. Organ. Stud. **41**(12), 1601–1625 (2020)
6. Currie, D.: Managing Employee Well-Being. Chandos Publishing (Oxford) Limited(2001)
7. Warr, P.: Psychology at Work. Penguin Group Books (2002)
8. Diener, E. (ed.): The Science of Well-Being, p. 37. Springer, Netherlands, Vol (2009)
9. Ryan, R.M., Deci, E.L.: On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. Annu. Rev. Psychol. **52**(1), 141–166 (2001)
10. Kashdan, T.B., Biswas-Diener, R., King, L.A.: Reconsidering happiness: the costs of distinguishing between hedonics and eudaimonia. J. Posit. Psychol. **3**(4), 219–233 (2008)
11. Christensen, J.O., Finne, L.B., Garde, A.H., Nielsen, M.B., Sørensen, K., Vleeshouwers, J.: The influence of digitalization and new technologies on psychosocial work environment and employee health: a literature review. Statens Arbeidsmiljøinstitutt, Nationalt Forskningscenter for Arbejdsmiljø (2020)
12. Adams, G., Salomons, T.V.: Attending work with chronic pain is associated with higher levels of psychosocial stress. Can. J. Pain **5**(1), 107–116 (2021)
13. Capecchi, S., Cappelli, C., Curtarelli, M., Di Iorio, F.: Synthetic indicators to analyze work-related physical and psychosocial risk factors: evidence from the European working conditions survey. Qual. Quant. **57**(6), 5713–5735 (2023)
14. Topp, C.W., Østergaard, S.D., Søndergaard, S., Bech, P.: The WHO-5 well-being index: a systematic review of the literature. Psychother. Psychosom. **84**(3), 167–176 (2015)
15. Jami, W.A., Kemmelmeier, M.: assessing well-being across space and time: measurement equivalence of the who-5 in 36 european countries and over 8 years. J. Well-being Assess. **4**, 419–445 (2020)
16. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

17. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
18. Hapfelmeier, A., Hothorn, T., Ulm, K., Strobl, C.: A new variable importance measure for random forests with missing data. Stat. Comput. **24**(1), 21–34 (2014)
19. Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S.: High-dimensional variable selection for survival data. J. Am. Stat. Assoc. **105**(489), 205–217 (2010)
20. Earle, D., Hurley, C.B.: Advances in dendrogram seriation for application to visualization. J. Comput. Graph. Stat. **24**, 1–25 (2015)
21. Hinsch, D.M., et al.: Associations between overcommitment, effort-reward imbalance and mental health: findings from a longitudinal study. Int. Arch. Occup. Environ. Health **92**(4), 559–567 (2018)
22. Vos, M., De Moortel, D., Vanroelen, C.: Effort-reward imbalance and the mental health of middle managers in Europe. European J. Public Health **32**(Suppl 3), ckac131.275 (2022)
23. De Sio, S., et al.: Telework and its effects on mental health during the COVID-19 lockdown. Eur. Rev. Med. Pharmacol. Sci. **25**(10), 3914–3922 (2021)
24. Haar, J.M., et al.: Outcomes of work-life balance on job satisfaction, life satisfaction and mental health: a study across seven cultures. J. Vocat. Behav. **85**(3), 361–373 (2014)
25. Russo, M.: Work hazards and workers' mental health: an investigation based on the fifth European working conditions survey. Med. Lav. **110**(2), 115–129 (2019)



Individual Causation with Additional Information

Giuseppe Demuru[1(✉)], Monica Musio[1], and A. Philip Dawid[2]

[1] Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
giuseppe.demuru2@unica.it

[2] Statistical Laboratory, University of Cambridge, Cambridge, England

Abstract. In many applications, such as medical disputes at Law, interest lies in whether an observed outcome in an individual case was caused by a specific exposure. An important approach aims to evaluate the probability that the outcome in this case was in fact caused by the exposure, using statistical data on similar individuals. Even the best possible experimental data on exposure and outcome typically can not identify this "probability of causation" (PC) exactly, but can only supply bounds for it. These bounds can be significantly improved by accounting for information about internal processes, using additional variables such as covariates and mediators. This work extends such bounds to the case where we have an instrumental variable.

Keywords: Causes of effects · individual causation · mediators · instrumental variable

1 Introduction

In causal inference it is important to distinguish two types of causal query which, although not entirely unconnected, are nevertheless very different, both in form and in the type of answer they require: Effects of Causes (EoC): 'I have a headache. I am wondering whether to take aspirin. Will that cause my headache to disappear?'; and Causes of Effects (CoE): 'I had a headache and took aspirin. My headache went away. Was that caused by the aspirin?' EoC type questions have the form 'Will A cause B ?' CoE questions have the form 'Did A cause B ?', and are also known as problems of individual causation. Thus the questions 'Does SARS-COV-2 vaccine cause thrombosis?', 'Ann is wondering whether to take a SARS-COV-2 vaccine, will that cause her to develop thrombosis?', are of EoC type; while the question 'Ann had a SARS-COV-2 vaccine and developed thrombosis. Was that caused by the vaccine?' is of CoE type.

The problem we address here is that of assessing "Causes of Effects": how might one use epidemiological data, gathered on groups, to assign a "probability of causation" (PC) in an individual case? This might be required in a court of law, where it is desired to establish responsibility for an adverse outcome: when

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 574–579, 2025.
https://doi.org/10.1007/978-3-031-64431-3_95

PC exceeds one half, a civil case might be regarded as proven on the "balance of probabilities". Thus suppose an individual, Ann, has developed a thrombosis, and we wish to investigate whether this might be due to the known fact that she had a prior SARS-COV-2 vaccine. We can consider PC as addressing the question: knowing that Ann was vaccinated against SARS-COV-2 and that she developed thrombosis, what is the probability that the (necessarily unobserved) potential response, which would have been observed had she not been vaccinated, would have been different? Statistical data are available on the exposures and outcomes of similar individuals. In previous work we showed that such data, however extensive, would not allow us to identify Ann's PC precisely. The data can, however, be used to set lower and upper bounds on PC. Although not as specific as one would ideally like, such bounds can be of use. In particular, if the lower bound exceeds one half then we can infer causality 'on the balance of probabilities'. We further showed that these bounds can be refined when it is possible to observe further variables in the data (though not necessarily on Ann) thus supplying partial information on the internal mechanistic workings of the causal 'black box'. Specifically, we considered pre-exposure variables (covariates), for example genetic predisposition, and variables intermediate between exposure and response (mediators), for example microclots, and assumed throughout that there was no confounding caused by still further unobserved variables. In this work we review previous results, and extend them to the case of an instrumental variable.

2 Probability of Causation and Its Bounds

For dealing with CoE questions we introduce the notion of "potential variables." If we have two variables, exposure X and outcome Y, the potential variable $Y(x)$ represents the value of Y that would occur if, either in reality or hypothetically, X were externatlly set to x (which we notate as $X \leftarrow x$). In the sequel we take both X and Y to be binary. We denote the pair $(Y(0), Y(1))$ by $\mathbf{Y}$: it is important to note that the pair $\mathbf{Y}$ cannot be fully observed. When X actually takes value x, we will observe $Y = Y(x)$; this is true in an interventional setting, by definition, and we assume ("consistency") that it is also true when the value of x is passively observed.

We assess CoE by means of the "probability of causation", PC, also known as the "probability of necessity", PN. If Ann took a SARS-COV-2 vaccine and subsequently developed thrombosis, the probability of causation in Ann's case is defined as:

$$\begin{aligned} \mathrm{PC}_A &= \mathrm{P}_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1) = \\ &= \Pr(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1) \\ &= \frac{\Pr(Y_A(0) = 0, Y_A(1) = 1 \mid X_A = 1)}{\Pr(Y_A(1) = 1 \mid X_A = 1)}. \end{aligned} \tag{1}$$

Here the suffix $_A$ relates to Ann's attributes.

Definition (1) involves the joint distribution of two variables related to the same subject, where one is observable while the other remains counterfactual. This makes PC_A typically unidentifiable. Nevertheless, we can establish bounds using the joint distribution of (X, Y) [1]. Suppose we have data on variables X and Y for individuals similar to (exchangeable with) Ann. Assuming no confounding between X and Y, expressed as independence between X and the pair of potential responses $\mathbf{Y} = (Y(0), Y(1))$, we can estimate probabilities like $\Pr(Y(1) = 1 \mid X \leftarrow 1) = \Pr(Y = 1 \mid X = 1)$ and $\Pr(Y(0) = 1 \mid X \leftarrow 0) = \Pr(Y = 1 \mid X = 0)$. This allows estimation of the margins of $(Y(0), Y(1))$, although we cannot estimate the full bivariate distribution of $(Y(0), Y(1))$, needed to compute (1). Nevertheless these margins constrain the joint probabilities, and we obtain the estimable inequality

$$\max\left(0, 1 - \frac{1}{\mathrm{RR}}\right) \leq \mathrm{PC}_A \leq \min\left(1, \frac{\Pr(Y = 0 \mid X \leftarrow 0)}{\Pr(Y = 1 \mid X \leftarrow 1)}\right). \tag{2}$$

Here, $\mathrm{RR} = \Pr(Y = 1 \mid X \leftarrow 1)/\Pr(Y = 1 \mid \mathrm{X} \leftarrow 0)$, termed the "experimental risk ratio". A $\mathrm{RR} > 2$ implies PC_A exceeds 50%, often crucial in legal proceedings. Refining these bounds is possible under appropriate assumptions, especially with measurable additional variables such as covariates.

2.1 Sufficient Covariate

We modify the initial condition by removing the "no confounding" assumption, introducing an additional "sufficient covariate", S. This covariate is measurable before exposure determination and can influence both exposure and outcome. We assume that, given S, there is no further confounding. This scenario is illustrated in Figure (1).

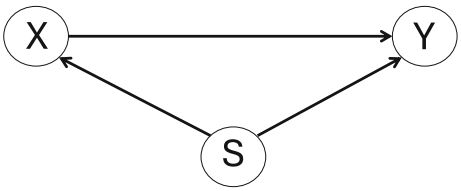


Fig. 1. Sufficient covariate

If we can observe S both in our dataset and for Ann, we can apply an analysis similar to that outlined above, but conditioned on Ann's specific value $S_A = s$. Instead, we assume that we can observe S for our dataset, so allowing estimation of the joint distribution of (X, S, Y), but do not observe S for Ann.

We introduce potential variables as follows:

$X(s)$**:** The potential exposure when S takes value s.

$Y(s, x)$**:** The potential response when S takes value s and X takes value x.

Given our assumption that there is no further confounding after conditioning on S, we can compute, for example, $\Pr(X(s) = x)$ as $\Pr(X = x \mid S \leftarrow s)$.

In this scenario, we obtain new bounds for PC_A [1]:

$$\frac{\Delta}{\Pr(Y = 1 \mid X \leftarrow 1)} \leq \mathrm{PC}_A \leq 1 - \frac{\Gamma}{\Pr(Y = 1 \mid X \leftarrow 1)}, \tag{3}$$

where

$$\begin{aligned}
\Delta &= \sum_s \Pr(S = s \mid X \leftarrow 1) \times \\
&\quad \max\{0, \Pr(Y = 1 \mid X \leftarrow 1, S \leftarrow s) - \Pr(Y = 1 \mid X \leftarrow 0, S \leftarrow s)\} \\
\Gamma &= \sum_s \Pr(S = s | X \leftarrow 1) \times \\
&\quad \max\{0, \Pr(Y = 1 \mid X \leftarrow 1, S = s) - \Pr(Y = 0 \mid X \leftarrow 0, S \leftarrow s)\}.
\end{aligned}$$

2.2 Mediators

In this Section we explore the bounds for PC_A in scenarios where a mediator plays a role in the causal pathway between the exposure X and the response Y. Specifically, we focus on cases where M serves as a complete mediator for the effect of X on Y (see Fig. 2).



Fig. 2. Complete mediator

We introduce potential variables: $M(x)$, representing the potential value of M when $X = x$, and $Y(m)$, representing the potential value of Y given $M = m$, with $M = M(X)$ and $Y = Y(M)$. Again we confine our analysis to the binary case and define $\mathbf{M} := (M(0), M(1))$. We assume exchangeability and non-confounding conditions, implying mutual independence between X, $\mathbf{M}$, and $\mathbf{Y}$.

Given experimental (or unconfounded observational) data providing probabilities

$$\Pr(M = m \mid X \leftarrow x) \quad \text{and} \quad \Pr(Y = y \mid M \leftarrow m),$$

we compute:

$$\Pr(Y = y \mid X \leftarrow x) = \sum_m \Pr(Y = y \mid M \leftarrow m) \Pr(M = m \mid X \leftarrow x). \tag{4}$$

When M is observed in the data but not for Ann, this additional evidence does not change the lower bound of 2, which disregards the presence of M. However we obtain an improved upper bound [3]:

$$\frac{\min\{ac + (1-d)(1-b),\ bc + (1-d)(1-a),\ ad + (1-c)(1-b),\ bd + (1-c)(1-a)\}}{(1-b)(1-c) + bd} \tag{5}$$

where

$$\begin{aligned} a &= \Pr(M = 0 \mid X \leftarrow 0) \\ b &= \Pr(M = 1 \mid X \leftarrow 1) \\ c &= \Pr(Y = 0 \mid M \leftarrow 0) \\ d &= \Pr(Y = 1 \mid M \leftarrow 1) \end{aligned}$$

The above analysis has been extended to the case where we have a direct mediation sequence $X = M_0 \rightarrow M_1 \rightarrow \cdots \rightarrow M_{n-1} \rightarrow M_n = Y$, and know the probabilistic structure of each link in the chain [2].

2.3 Instrumental Variable

In this Section we consider the case where we have an instrumental variable. This involves, in addition to the treatment variable X and response variable Y, a further observed variable Z (the instrument), and a generally unobserved variable U. Again we take X and Y as binary; U is typically multivariate.

We suppose (see Fig. 3):

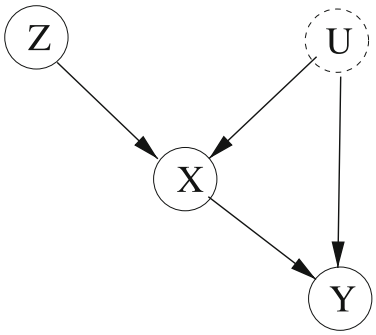


Fig. 3. Instrumental variable

1. U is a set of pre-existing characteristics of the individual
2. Z is associated with X, but not with U
3. Given X, and the individual characteristics U, the response Y is unaffected by Z.

We also suppose no further confounding.

For Ann, we have observed $Z = z$, $X = 1$ and $Y = 1$, and are interested in PC_A, the probability that it was X that caused Y. We have data on Z, X and Y, and possibly also U, from which it is possible to estimate their underlying probabilistic structure. The data may be observational, or involve direct intervention on X.

In the (exceptional) scenarios where the full probabilistic structure for all variables (including U) is known, the information on Z is of no relevance to PC_A, and we are back to the case of a sufficient covariate U observed in the data, as in Sect. 2.1.

When, as will be more typical, we don't observe U in the observational data, this leads to unmeasured confounding between X and Y; disregarding this yields wrong conclusions [4]. In order to proceed, we require both experimental data, where X is intervened on, and observational data, on Z, X and Y. We then have

$$l(z) \leq \mathrm{PC}_A \leq u(z)$$

where, from equation (25) of [5] applied conditionally on $Z = z$,

$$l(z) = \max\left\{0, \frac{\Pr(Y=1 \mid Z=z) - \Pr(Y=1 \mid X \leftarrow 0, Z=z)}{\Pr(X=1, Y=1 \mid Z=z)}\right\}$$
$$u(z) = \min\left\{1, \frac{\Pr(Y=0 \mid X \leftarrow 0, Z=z) - \Pr(X=0, Y=0 \mid Z=z)}{\Pr(X=1, Y=1 \mid Z=z)}\right\}.$$

3 Discussion

Even with significant unconfounded experimental data that precisely elucidate the probabilistic dependence of a response variable on an exposure variable, we usually can only set bounds for the probability of causation when an individual develops the response after exposure. However, these bounds can be improved if we have supplementary information, such as data on covariate, mediator, or instrumental variables. Here we have examined relatively straightforward cases with additional information; more complex scenarios and real-world applications require further investigation.

References

1. Dawid, A.P., Musio, M., Murtas, R.: The probability of causation. Law Probab. Risk **16**(4), 163–179 (2017)
2. Dawid, A.P., Humphreys, M., Musio, M.: Bounding causes of effects with mediators. Sociol. Methods Res. **53**(1), 28–56 (2024)
3. Dawid, A.P., Murtas, R., Musio, M.: Bounding the probability of causation in mediation analysis. In: Topics on Methodological and Applied Statistical Inference (2016)
4. Musio, M., Dawid, A.P.: Individual causation with biased data. arXiv:2311.08242 (2024)
5. Tian, J., Pearl, J.: Probabilities of causation: bounds and identification. Ann. Math. Artif. Intell. **28**, 287–313 (2000)



Indexing Social Unfitness
Quantifying Ordinal Levels by Problems Severity

Giulio D'Epifanio(✉)

Department of Political Science,University of the Study of Perugia,Perugia, Italy
ggiuliodd@gmail.com

Abstract. In order to help a Policy-Maker(PM) in comparing districts (space or time related social domains) with regard to social-unfitness (related to eg health deficiency, dissatisfaction in daily life, psychological discomfort, etc.), a disvalue-functional is used which provides a social index. Constructed on a problems-checklist, ordered by increasing severity, such an index is standardized on PM's specifications. The underlying methodological question concerns the operative, possibly automatized, way for assigning meaningful standardized scores to parameters increment of dis-value, related to fitness-decaying, which graduate the severity-levels on the scale which has been associated to the problems-checklist. A principle of intrinsic loss-of-value, related to fitness decaying on the problems-checklist, is outlined according to the which the current fitness-status value of an individual is associated to his (properly standardized) capacity of resisting to insurgences of further fitness problems. Applying such a principle on a probability framework (standardized conditionally on the PM's specifications), the increments of dis-value (which will enter index) are interpreted as non-transition probabilities on the problems-checklist. These parameters could be operatively scored (by using model-based data-analysis tools) calibrated on training-data, associated to the reference social-behavior (eg a social good practice) chosen by the PM

Keywords: social indexing · scale quantifying · unfitness · health-deficiency

1 Problems-Based Scaling

The PM would to compare[1] situations of districts $\{A_1, A_2, \ldots, A_q\}$, with regard to social unfitness, using districts data as eg those in table 1 below[2].

The methodological question herein concerns a structured operative, possibly automatized, way for assigning a (properly standardized) meaningful value to

[1] for general references on evaluations, eg see[1] [5].

[2] occurrences of problems have been counted herein on various districts; an ordinal scale Y is defined which classifies individuals on unfitness levels $l := 0, 1, 2, 3, 4, 5, 6$, which could be re-labeled, for example as: "*excellent*", "*good*","*more than acceptable*","*acceptable*","*not acceptable enough*","*bad*","*very bad*".

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 580–585, 2025.
https://doi.org/10.1007/978-3-031-64431-3_96

Table 1. Example data

level of unfitness Y:	0	1	2	3	4	5	6
district:							
A1	0	16	24	31	12	0	4
A2	4	28	65	107	26	1	3
A3	4	42	71	102	33	0	2
A4	4	71	112	194	53	0	4

(a) Data by the districts to be benchmarked

Y:	0	1	2	3	4	5	6
reference district A^*	12	157	272	434	124	1	13

(b) Data by the reference district A^*

parameters "increment of dis-value" ω_l, defined in (2) below on an assigned problems-checklist (1), in order to graduate the levels of severity on ordinal scale (eg see [8]) Y and, consequently, to construct a dis-value index (3) to be applied to the various districts of interest. In advancing severity along problems-checklist (1), a principle of "intrinsic loss-of-value" is outlined, re-interpreting a similar criterion used in [4]. Such a principle is then reformulated standardized on a probability referenced-setting on the which the "increments of dis-value" ω_l (which will enter social index) could be operatively extracted by training-data, eg those coming from the district A^* which has been chosen to represent the reference social behavior.

1.1 Severity Ordinal Scale Setting

Suppose that a problems-checklist, designed to be used on any individual, in a specified domain-type[3] $\mathcal{D}$, has been provided by experts to check occurrence[4] (yes/not) of increasingly more severe problems

$$P_0 \preceq P_1 \preceq P_2 \preceq ... \preceq P_l \preceq ... \preceq \ldots \preceq P_{L-1} \preceq P_L := P_{Full} \tag{1}$$

Here, P_0 represents the base-line "dummy problem" which always will occurred by anyone (even when "no actual problem occurs", conventionally detected by a "dummy" diagnostic sensor), P_{Full} represents the most severe conceived problem; $P_l \preceq P_{l+1} \Leftrightarrow$ "*whenever problem* P_{l+1} *occurs, also problem* P_l *has been occurred*" (Guttman-like ordering). Proposed by experts, chain (1) represents a path in fitness-decaying, from the ideal "better fitness status" (where "no actual problems occurs") toward the worst one (where "the most severe conceived problem occurs"). Any individual in $\mathcal{D}$, checked against chain (1), will be classified on unfitness severity-level $(l-1)$ whenever "problem P_{l-1} is detected but not also the next problem P_l"; so that a severity ordinal scale

[3] eg "18 to 25 year old, female" perhaps with a certain "social background".

[4] eg in evaluating "health deficiency", a sequence of binary-outcome clinical-diagnostic sensors could be implemented; each sensor emits a binary signal, perhaps by integrating various signals from more basic clinical sensors.

$Y \in \{0, 1, ..., l, ..., L\}$ remains operatively defined associated to chain (1) and vice versa[5] ie $P_l \leftrightarrow (Y \geq l)$, $l := 0, \ldots, L$.

1.2 Pro-capita Standardized Disvalue in Fitness Decaying

Consider a $\mathcal{D}$-typiced individual from the reference-district A^* which has been chosen by the PM for standardization. Sequentially checking such an individual on problems-chain (1), suppose that it transits[6] from the $(l-1)$-th to the l-th level of severity scale Y. Suppose then that, related to such a fitness-decaying transition, experts have been able to specify " increment of dis-value" :

$$\omega_l := \Delta_{l-1} disV := disV(P_l) - disV(P_{l-1}) \geq 0 \tag{2}$$

ie the "value loss", ("pro-capita" -)standardized[7] on A^*. Here, $disV(.)$ denotes a (non negative, not decreasing) dis-value function such that $disV(P_l) \geq disV(P_{l-1}) \geq 0$ $(l := 1, \ldots, L)$, to be conventionally initialized so that $disV(P_0) := 0$ (the "dummy increment" being let $\omega_0 := 0$).

1.3 Indexing Social Unfitness

Consider any actual district $A \in \{A_1, A_2, \ldots, A_q\}$, with regard to the insurgency of problem P_l on checklist (1). Then, the product

$$(disV(P_l) - disV(P_{l-1})) \cdot RelFreq\{Y > l-1\}[A]$$

would represent the "*increment of social dis-value*" (standardized on A^*) due to the percentage-mass $RelFreq\{Y > l-1\}[A] \cdot 100\%$ of the individuals, in district A, which transit (since problem P_l has insurged on them) from the $(l-1)$-th to the l-th level of severity scale Y. By integrating such social dis-value increments, $l := 1, \ldots, L$, it follows dis-value functional[8]:

$$A :\mapsto W[A] := \sum_{l:=1}^{L} \omega_l \cdot (1 - F_Y[p[A]](l-1)) = \sum_{l:=1}^{L} s(l;\, \omega) \cdot p_l[A] \tag{3}$$

Here[9], $p[A] := (p_0, p_1, \ldots, p_L)[A]$ denotes the relative distribution which describes the situation of district A on unfitness scale Y, $F_Y[p(A)](l-1) :=$

[5] there is the logical identification of proposition *"problem P_l is occurred"* with that of "*the outcome-level of Y is at least l*".

[6] due to insurging problem P_l given that less severe problems $P_1, \ldots, P_{l-1}$ have already occurred.

[7] e.g. in evaluating "health deficiency" , experts could evaluate the added "social-economic dis-value" , conventionally re-proportioned pro-capita, due to the health declining of (say) 100 typical individuals in reference district A^* on the which problem P_l has insurged, being problems $P_1, \ldots, P_{l-1}$ already occurred.

[8] such an operator (provided that parameters ω_l $(l := 1, \ldots, L)$ have been specified), applied to any actual district A, takes into input its actual data to assign A^*-standardized score $W[A]$ in order to position A against the concurrent districts.

[9] $RelFreq\{Y > l-1\}[A] = RelFreq\{Y \geq l\}[A] = p_l(A) + p_{l+1}(A) + \cdots + p_L(A) = 1 - (p_0(A) + p_1(A) + \cdots + p_{l-1}(A)) = 1 - RelFreq\{Y \leq l-1\}[A] = 1 - F_Y[p(A)](l-1)$.

$p_0(A) + p_1(A) + \cdots + p_{l-1}(A)$ the cumulative distribution; $s(l;\, \omega) := \omega_1 + \omega_2 + \cdots + \omega_l$ denotes the quantifier of dis-value level l $(l := 1, \ldots, L)$ on Y. Functional[10] $W[.]$ provides an index to be operativized by adopting some criterion to determine parameters $\omega = (\omega_1, .., \omega_L)$.

2 Specifying *social Dis-Value* on a probabilistic setting

Let $\mathcal{F}^*$ denote a certain specified reference-framework which includes the choice of standard district A^*. It could be assumed the principle that[11]:

"the bigger the $\mathcal{F}^*$-standardized expected resistance[12] against insurgency of more severe problem P_l, given that problem P_{l-1} has occurred; the bigger the $\mathcal{F}^*$-standardized intrinsic "loss of value" in the current fitness-status[13] of that any real individual on the which problem P_l actually insurges". Then, by interpreting such a principle on a probabilistic setting, the A^*-standardized "*propensity in resisting to further decaying*" from level $(l-1)$ could be represented by "non-transition probability" $Pr\{Y = l-1 | Y \geq l-1; \mathcal{A}^*\}$ on chain (1), $l := 1, \ldots, L$. Here, $\mathcal{A}^*$ represents the population of the individuals in the district A^*whose social behavior has been chosen by the PM as the standard one. Thus, "dis-value increases" (2) could be now formally associated (up to some monotone transformation[14] $\varphi_l(.)$) to parameters

$$\omega_l^* := \varphi_l(Pr\{Y = l-1 | Y \geq l-1; \mathcal{A}^*\}),\ l := 1, \ldots L \tag{4}$$

to be[15] estimated[16]

[10] Incidentally, such an index could be viewed as an instance of a more general class of (Yaari-Quiggin-like) value-functionals in "utility theory" (eg see [2]), so inheriting formal properties of rationality in social choices from the RDEU theory (e.g. see [2], pp. 559).

[11] the current fitness-status value of an individual is related to his ($\mathcal{F}^*$-standardized) "capacity in resisting" to insurgences of further more severe problems; a criterion of "intrinsic worthiness" (related to goal-achievements) is reinterpreted reversed here, see [3][4].

[12] ie the expected resistance, for a typical individual in A^*, to further fitness-decaying beyond the $(l-1)$th severity level.

[13] eg the intrinsic value of its health-status is related to his (properly standardized) inability in resisting to further health decaying; for reference to health related concepts eg see [7].

[14] $\varphi_l(.)$ could specify design-requirements (eg to represent losses due to types of consequences related to transitions); eg it might be the identity function, or a transformation which assures "additivity" on chain (1), see [3][4].

[15] of course, $Pr\{Y=l-1|Y\geq l-1;\mathcal{A}^*\} = \frac{Pr\{Y=l-1;\mathcal{A}^*\}}{Pr\{Y\geq l-1;\mathcal{A}^*\}} = \frac{p^*_{l-1}}{p^*_{l-1}+p^*_l+\cdots+p^*_L}$.

[16] for instance, recalling data-table (1b): $\omega_0 = 0$, $\omega_1 = Pr\{P_1\ fails\} \simeq \frac{12}{12+157+272+434+124+1+13}$ $\omega_2 = Pr\{P_2\ fails | P_1\ detected\} \simeq \frac{157}{157+272+434+124+1+13}$, $\omega_3 = Pr\{P_3\ fails | P_2\ detected\} \simeq \frac{272}{272+434+124+1+13}$, $\omega_4 \simeq Pr\{P_4\ fails | P_3\ detected\} = \frac{434}{434+124+1+13}$, $\omega_5 = Pr\{P_5\ fails | P_4\ detected\} \simeq \frac{124}{124+1+13}$, $\omega_6 \simeq Pr\{P_6\ fails | P_5\ detected\} = \frac{1}{1+13}$; then (let here $\varphi_l(.)$ be the iden-

from data of A^*; thus, specializing index (3) to

$$A :\longmapsto W[A;\, \omega^*] := \sum_{l:=1}^{L} \varphi_l\left(\frac{Pr\{Y=l-1;\mathcal{A}^*\}}{Pr\{Y\geq l-1;\mathcal{A}^*\}}\right) \cdot (1 - F_Y[p[A]](l-1)) \qquad (5)$$

possibly re-ranging[17] meaningfully it on interval $[0, 100\%]$:

$$A :\longmapsto W^*[A;\, \omega^*] := \frac{W[A;\, \omega^*] - W[A_{best};\, \omega^*]}{W[A_{worst};\, \omega^*] - W[A_{best};\, \omega^*]}$$
$$= \sum_{l:=1}^{L} \frac{\omega_l^*}{\sum_{r:=1}^{L} \omega_r^*} \cdot (1 - F_Y[p[A]](l-1))$$

2.1 Model-Based Multi-domain Indexing

To take into account various *"reference domains"* $\mathcal{D}_{|x_r}$, related to specific conditions $x := X \in \{x_1, \ldots, x_r, \ldots, x_R\}$ of individuals (eg stratified by sex, age and social status, etc.), index (5) could be adapted as it follows:

$$A \longmapsto \sum_{r:=1}^{R} q_r \cdot \{ \sum_{l:=1}^{L} \varphi_l\left(\frac{\exp(\widehat{a}_l + \hat{b}_l \cdot x_r)}{1 + \exp(\hat{a}_l + \hat{b}_l \cdot x_r)}\right) \cdot (1 - F_{Y|x_r}[p_{|x_r}[A]](l-1)) \}$$

Here, a sequence of logistic models (by crossing levels $l := 1, \ldots, L$ on chain (1)) is used to assign value-increases $\omega^*_{l|x_r}$, which are $\mathcal{D}_{|x_r}$-domain specific. Parameters $\hat{a}_l$ and $\hat{b}_l$ could be determined from (real o perhaps virtual) training-data associated to the standard reference-district A^*. Here, $q_r \geq 0$ represents weights ($\sum_{i:=1}^{R} q_r = 1$) which could meaningfully interpret the political relevance of the various "social reference domains", perhaps in the perception of a certain stakeholder. A complex example, which uses structured modeling to determine $\omega^*_{l|x_r}$, is presented in [4]. In multi-domain indexing, it could be of interest to consider (in a simulation heuristic study) effects due to different choices about the standard district A^*, thus interpreting various (real o perhaps virtual) social behaviors to be considered as "reasonable standard" .

tity), by accumulating dis-value-increases yields scale quantification:

	0	1	2	3	4	5	6
level-score of Y:	0	0.011846	0.168689	0.490964	1.249705	2.148256	2.219685

.

[17] the dis-value $W[A]$, due to social unfitness, is considered here compared against the extreme virtual ideal situations: the "best" social situation A_{best} (such that $p[A_{best}] := (100\%, 0, \ldots, 0)$ so that $W[A_{best};\, \omega^*] = 0$) and the "worst" one A_{worst} (such that $p[A_{worst}] := (0, 0, \ldots, 0, 100\%)$ so that $W[A_{worst};\, \omega^*] = \omega_1^* + \cdots + \omega_L^*$)).

References

1. Bird, S.M., Cox, D., Farewell, V.T., Goldstein, H., Holt, T., Smith, P.C.: Performance indicators: good, bad, and ugly. JRSS-A **168**, 1–27 (2005)
2. Chateauneuf A., Cohen M., Meilijson I.: Four notions of mean-preserving increase in risk attitudes and applications to the rank-dependent expected utility model. J. Math. Econ. **40**, 547-571¡hi rend="it" /¿ (2004)
3. D'Epifanio ,G.: Sviluppo di un Indice Multi-attributo per la Valutazione del Merito. In: Criteri e indicatori per misurare l'efficacia delle attività universitarie, vol I, pp. 279, CLEUP, Padova (2011)
4. D'Epifanio, G.: Indexing the normalized worthiness of social agents. In: Perna, C., et al. (eds.) Springer Proceedings in Mathematics & Statistics 227, pp. 263–274. Springer International Publishing AG, Cham, Switzerland, Studies in Theoretical and Applied Statistics (2018)
5. Fisher, N.I.: A comprehensive approach to problems of performance measurement. JRSS-A **182**(3), 755–803 (2019)
6. Franco-Santos, M., Lucianetti, L., Bourne, M.: Contemporary performance measurement system: a review of their consequences and a framework for research. Manag. Account. Res. **23**(2), 79–119 (2012)
7. Hawthorne, G., Richardson, J., Osborne, R.: The assessment of quality of life (AQoL) instrument: a psychometric measure of health-related quality of life. Qual. Life Res. **8**(4), 209–224 (1999)
8. Kampen, J., Swyngedouw, M.: The ordinal controversy revisited. Qual. Quant. **34**(1), 87–102 (2000)

A Markov-Switching Model for Studying Extreme Temperatures in *Friuli Venezia Giulia*, Italy

Gioia Di Credico(✉), Vincenzo Gioia, and Francesco Pauli

Department of Economic, Business, Mathematical and Statistical Sciences, "Bruno de Finetti", University of Trieste, Trieste, Italy
gioia.dicredico@deams.units.it

Abstract. Heat waves, generally defined as prolonged periods of extreme temperatures, have hit several areas of the world and have been observed more frequently in recent decades. Several proposals appeared in the literature for classifying heat waves, mainly related to expert-based and area-specific fixed thresholds or quantile-based climate-model approaches. Summer temperature patterns exhibit different stochastic processes, which can be determined through a latent variable that describes the membership of each observation to a normal or a heat wave regime. We accommodate these characteristics by exploiting a Markov-switching Bayesian additive model for tail behavior modeling and aiming at the probabilistic classification of the heat wave regime. We illustrate the proposal by analyzing the maximum daily temperatures in four locations of the Italian region *Friuli Venezia Giulia*.

Keywords: Bayesian additive model · heat waves · regime-switching

1 Introduction

Climatological studies are becoming more relevant for supporting the challenges posed by the XXI century's public agenda [1]. Heat waves are one of the most relevant aspects of climate due to the potential effect of prolonged high temperatures on ecological systems and human health [3,6]. Heat waves are loosely defined as extended periods during which the temperature stays above a threshold. The precise definition to adopt in a study also depends on the type of impact, which is primarily of interest [7]. Upon adopting a specific definition, examining the frequency, intensity, and duration of heat waves is pertinent. However, this task is made difficult by the limited availability of data, which results from the extreme characteristics of heat waves and their varied occurrence across different regions [8]. Motivated by this practical challenge, we propose a flexible and interpretable approach enhancing the probabilistic classification of extreme summer temperatures, that leverages a multi-parameter Markov-switching Bayesian additive model focusing on tail behavior [10,11]. The proposal is illustrated by analyzing the maximum daily temperatures recorded in some meteorological stations of the Italian region *Friuli Venezia Giulia* (FVG).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 586–591, 2025.
https://doi.org/10.1007/978-3-031-64431-3_97

2 Markov-Switching for Extreme Value Modeling

Markov-switching regression models account for different stochastic processes along different periods by considering a latent variable controlling which regime is applied to each observation. Let y_t, $t = 1, \ldots, T$, be the realizations of an observable stochastic process Y_t and S_t, $S_t \in \{1, 2\}$, a two-state hidden process ruled by a first-order Markov-chain structure. We assume that the Y_t's are independent conditionally to S_t denoting their conditional distribution with $Y_t|S_t = s_t \sim \mathcal{D}_{s_t}(\boldsymbol{\theta}^t_{s_t})$, where $\mathcal{D}_{s_t}(\cdot)$ is a parametric distribution family characterized by a q_{s_t}-dimensional parameter $\boldsymbol{\theta}^t_{s_t}$.

In the following, we consider a model where the first regime is Gaussian and the second one can follow a Gaussian or a generalized extreme value (GEV) distribution. In this respect, we link the regime-specific distributional parameter $\theta^t_{s_t r}$ to the linear predictor components $\eta^t_{s_t r}$ via

$$g_{s_t r}(\theta^t_{s_t r}) = \eta^t_{s_t r} = \sum_k h_{k s_t r}(x^t_k), \quad r = 1, \ldots, q_{s_t}, \tag{1}$$

where $g_{s_t r}(\cdot)$ is a suitable link function and $h_{k s_t r}(\cdot)$ can be a linear or smooth function of the covariate x^t_k.

The latent process S_t considers a two-state discrete-time Markov chain. We denote with $\boldsymbol{\pi}^t = (\pi^t_1, \pi^t_2)^\top$ the probability vector of being in the two states at time t. Here, $\boldsymbol{\pi}^1 = (\Pr(S_1 = 1), \Pr(S_1 = 2))^\top$, while, for $t = 2, \ldots, T$, $\pi^t_{s_t} = \Pr(S_t = s_t|\Omega_t, \boldsymbol{\theta}^t_{s_t})$, with Ω_t including all the information until time t. Finally, the transition probabilities are specified by the stochastic matrix $\boldsymbol{\delta}^t$, where $\delta^t_{jk} = \Pr(S_t = k|S_{t-1} = j)$, for $j, k = 1, 2$, with $j \neq k$.

By considering a linear-basis expansion, the linear predictor components in (1) can be expressed as $\eta^t_{s_t r} = \mathbf{x}^t_{s_t r}\boldsymbol{\beta}_{s_t r}$, where $\mathbf{x}^t_{s_t r}$ is a $p_{s_t r}$-dimensional row vector of covariate and $\boldsymbol{\beta}_{s_t r}$ is the associated parameter. The switching state probabilities, δ^t_{12} and δ^t_{21}, can be modeled in logit scale and could vary as a function of the covariates; hence, we denote the related parameters with $\boldsymbol{\beta}^\top_{\delta_{12}}$ and $\boldsymbol{\beta}^\top_{\delta_{21}}$. According to the Hamilton filter [4], the log-likelihood for $\boldsymbol{\beta} = (\boldsymbol{\beta}^\top_{11}, \boldsymbol{\beta}^\top_{21}, \boldsymbol{\beta}^\top_{12}, \boldsymbol{\beta}^\top_{22}, \ldots, \boldsymbol{\beta}^\top_{\delta_{12}}, \boldsymbol{\beta}^\top_{\delta_{21}})^\top$ is

$$\ell(\boldsymbol{\beta}) = \sum_{t=2}^{T} \log f_{Y_t}(y_t|\Omega_t, \boldsymbol{\beta}) = \sum_{t=2}^{T} \log \left\{ \sum_{j,k=1}^{2} \delta^t_{jk} \pi^{t-1}_j f_{Y_t|S_t}(y_t|s_t = k, \Omega_t, \boldsymbol{\beta}) \right\},$$

where $\pi^t_j = \sum_{k=1}^{2} \delta^t_{kj} \pi^{t-1}_k f_{Y_t|S_t}(y_t|s_t = k, \Omega_t, \boldsymbol{\beta}) / f_{Y_t}(y_t|\Omega_t, \boldsymbol{\beta})$. Whenever feasible, we select weakly informative priors scaled to the data for the intercepts and linear effects. Alternatively, we set weakly informative normal priors scaled according to frequentist inference results or following the literature. The degree of regularization on the smooth terms is ruled by half student-t priors on the inverse of the smoothing parameter (see, e.g., [11]). We perform Bayesian inference employing the No-U-Turn Sampler extension to the Hamiltonian Monte Carlo, implemented in the probabilistic programming language `Stan`.

3 Summer Temperatures in *Friuli Venezia Giulia*, Italy

The Regional Environmental Protection Agency (Arpa) of the Italian Region FVG provides historical weather data for several meteorological stations, at hourly resolution and starting from 1991 [9]. We consider four stations reflecting the heterogeneous territorial conformation of this region: `Pordenone`, in the lowlands of the Po-Venetian Valley; `Trieste`, lying on the homonym gulf and belonging to the Karst region; `Capriva` (del Friuli), located in the hilly Collio region and close to Gorizia; `Zoncolan`, in the Carnic Alps. Figure 1 shows the position and the altitude of the four stations in the FVG region.

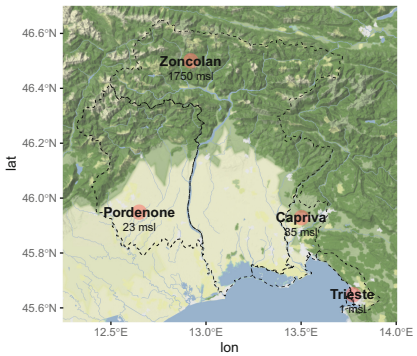


Fig. 1. Geographical map of the FVG region and the four meteorological stations.

For each of the four sites, we analyze the maximum daily temperature (in Celsius) from May to September, spanning 1995 - 2023. We removed the few missing values due to maintenance periods or temporary malfunctions. Thus, the number of data points T is 4354 for `Capriva`, 4378 for `Pordenone`, 4404 for `Trieste`, and 4281 for `Zoncolan`. We consider as predictors the year (`year`) and the day of the year (`doy`), thus accounting for trend and seasonality patterns.

We specify a model where the first regime R1 is Gaussian and the second one R2 may be one among Gaussian (labeled as `Gaussian - Gaussian`), Gumbel (`Gaussian - Gumbel`), or three parameters GEV (`Gaussian - GEV`). In the following, we omit the station-specific index to lighten the notation. For all the considered regime distributions, we model the location parameters $\mu^t_{s_t}$ as

$$\mu^t_{s_t} = \eta^t_{s_t 1} = h_{1 s_t 1}(\texttt{year}_t) + h_{2 s_t 1}(\texttt{doy}_t)\,, \quad t = 1, \ldots, T,\ s_t = 1, 2, \tag{2}$$

where $h_{1 s_t 1}$ and $h_{2 s_t 1}$ are, respectively, a linear and a smooth function. The log-scale parameter $\log \sigma^t_{s_t} = \eta^t_{s_t 2} = \beta_{s_t 2}$, the shape parameter $\xi^t_{s_t} = \eta^t_{s_t 3} = \beta_{s_t 3}$ of the GEV regime ($s_t = 2$), and the logit switching probabilities are modeled via intercepts. The posterior distribution of the shape parameter in the `Gaussian - GEV` model concentrates on negative values, entailing the reverse

Weibull distribution. Unreported results have shown no benefit of modeling the other distributional parameters using covariates.

For each station, we compare the different parametric specifications of the two regimes using the leave-one-out information criterion (LOOIC). The results in Table 1 highlight that the `Gaussian - GEV` model outperforms the `Gaussian - Gaussian` and `Gaussian - Gumbel` models. However, analyzing the differences in the expected log posterior density and the associated standard error shows weak evidence of the gain achieved by considering the `Gaussian - GEV` model for `Zoncolan`. The setting `Gaussian - Gumbel` outperforms `Gaussian - Gaussian` only in the `Trieste` case.

Table 1. LOOIC. The underline indicates the best model for each location.

Station/Model	`Gaussian - Gaussian`	`Gaussian - Gumbel`	`Gaussian - GEV`
`Capriva`	21830.9	21840.8	<u>21788.9</u>
`Pordenone`	21141.5	21176.1	<u>21112.6</u>
`Trieste`	19455.8	19371.6	<u>19348.3</u>
`Zoncolan`	21708.0	21795.1	<u>21706.7</u>

Analyzing the results of the fitted trend in the `Gaussian - GEV` models, we observe that: (i) the effect is estimated to be positive for all the stations; (ii) and there are no differences in the point estimates for the two regimes of each station, except for `Capriva`. More precisely, the posterior median estimates of the trend with 95% credible interval (in brackets) are 0.05 for R1 and 0.04 for R2 (R1: 0.04 - 0.07, R2: 0.02 - 0.06) for `Capriva`; 0.07 (R1 and R2: 0.06 - 0.09) for `Pordenone`; 0.03 (R1: 0.01 - 0.04, R2: 0.02 - 0.05) for `Trieste`; 0.05 (R1: 0.03 - 0.06, R2: 0.02 - 0.06) for `Zoncolan`. The slight differences in the trend estimates could reflect some aspects of the territorial conformation. Indeed, the higher estimate is obtained for the lowland `Pordenone`, while the lower one is for `Trieste`, where a mitigation effect is probably due to wind currents and the sea proximity. About the seasonality effect, the maximum daily temperature peak is estimated to be between July and August for both regimes. Figure 2 reports the time series of the observed maximum daily summer temperature for the four selected stations, considering 2023. Several European countries, including Italy, experienced repeated heat waves that year. Considering the posterior deseasonalised estimated mean effect thresholds, the *Cerberus* and *Charon* heat waves occurring in the second and third week of July are identifiable with high confidence as heat waves by the proposed approach. Moreover, the African anticyclone, occurred during the third week of August and characterized by even more extreme temperatures, is also associated with high posterior probabilities of being in the heat wave regime.

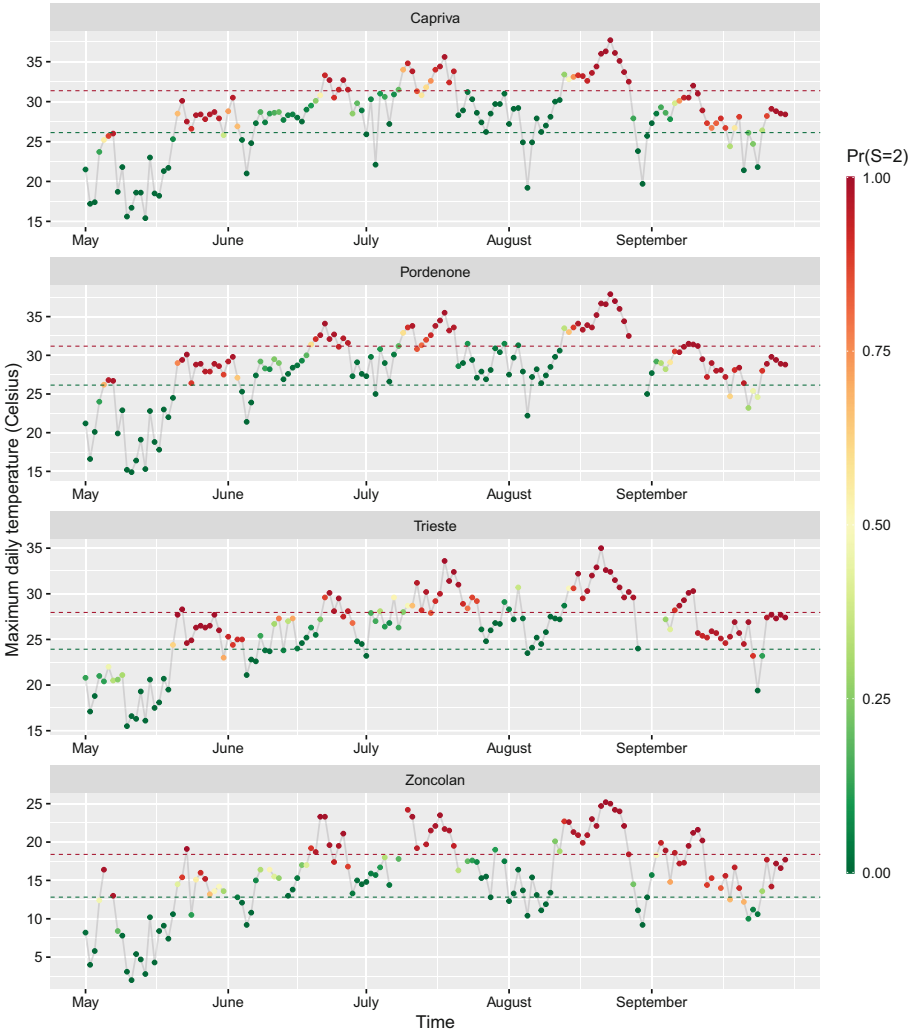


Fig. 2. Time series of the observed maximum daily summer temperature for all the stations and `year`=2023. The color gradient scale reports the probability of being in a heat wave regime. The dashed lines report the deseasonalized mean effect for both regimes. Note the different scales on the y-axis. Gaps rely on missing observations.

4 Conclusions

The periods associated with high probability of being in the extreme temperature regime from the proposed model are in line with those of the well-known European summer 2023 heat waves. Furthermore, despite the different geographical characteristics of the four stations, we obtain similar posterior probability estimates of being in the heat wave regime, positive trends and akin seasonality effects across

sites. Although we account only for relevant calendar variables in location parameter modeling, site-specific weather-related observations (or forecasts) and large-scale climatic indices could be included to model the regime-specific distributional parameters and the transition probabilities. This exploratory study is promising for spatial modeling extensions. In this respect, copula-based approaches can be considered for several purposes: (i) coupling maximum and medium temperature into a joint marginal Markov-switching with dynamic bivariate copula regression modeling [5]; (ii) clustering similar small climatic areas for identifying representative meteorological stations [2]; (iii) capturing time-space evolution in more extended geographical domains [12].

References

1. Agenda 21.: United Nations Conference on Environment & Development, Rio de Janeiro, Brazil, 3 to 14 June 1992. Available at https://sustainabledevelopment.un.org/content/documents/Agenda21.pdf
2. Benevento, A., Durante, F., Pappadà, R.: An approach to cluster time series extremes with spatial constraints. In: Chelli, F.M., Ciommi, M., Ingrassia, S., Mariani, F., Recchioni , M.C. (eds.), Book Short Pap. SIS 2023, 679–684 (2023)
3. Forzieri, G., Cescatti, A., Silva, F.B., Feyen, L.: Increasing risk over time of weather-related hazards to the European population: a data-driven prognostic study. Lancet Planet. Health **1**(5), e200–e208 (2017)
4. Hamilton, J.D.: Regime switching models. In: Durlauf, S.N., Blume, L.E. (eds.) Macroeconometrics and Time Series Analysis. TNPEC, pp. 202–209. Palgrave Macmillan UK, London (2010). https://doi.org/10.1057/9780230280830_23
5. Marra, G., Radice, R.: Bivariate copula additive models for location, scale and shape. Comput. Stat. Data Anal. **112**, 99–113 (2017)
6. Pauli, F., Rizzi, L.: Statistical analysis of temperature impact on daily hospital admissions: Analysis of data from Udine, Italy. Environ. Off. J. Int. Environ. Soc. **17** (1), 47–64 (2006)
7. Perkins, S.E.: A review on the scientific understanding of heatwaves–their measurement, driving mechanisms, and changes at the global scale. Atmospheric Res. **164–165**, 242–267 (2015)
8. Perkins-Kirkpatrick, S.E., Lewis, S.C.: Increasing trends in regional heatwaves. Nat. Comm. **11**, 3357 (2020)
9. Regional Meteorological Service Arpa Friuli Venezia Giulia www.arpa.fvg.it/
10. Shaby, B.A., Reich, B.J., Cooley, D., Kaufman, C.G.: A Markov-switching model for heat waves. Ann. Appl. Stat. **10**(1), 74–93 (2016)
11. Umlauf, N., Klein, N., Zeileis, A.: BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). J. Comp. Graph. Stat. **27**(3), 612–627 (2018)
12. Winter, H.C., Tawn, J.A., Brown, S.J.: Modelling the effect of the El Nino-Southern Oscillation on extreme spatial temperature events over Australia. Ann. Appl. Stat. **10**(4), 2075–2101 (2016)



Estimating Cure Fraction in Population-Based Survival Colon Cancer Data

Fabrizio Di Mari[1(✉)], Roberto Rocci[1], Giovanna Tagliabue[2], Silvia Rossi[3], and Roberta De Angelis[3]

[1] Sapienza Università di Roma, Rome, Italy
{fabrizio.dimari,roberto.rocci}@uniroma1.it
[2] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
giovanna.tagliabue@istitutotumori.mi.it
[3] Istituto Superiore di Sanità, Rome, Italy
{silvia.rossi,roberta.deangelis}@iss.it

Abstract. In recent decades, cancer prognosis has notably improved, leading to a higher rate of patient cure. Monitoring the proportion of cured individuals is valuable for patients and provides insight into survival trends for treatable diseases. It is crucial to estimate cure fraction in a diseased population when other competing moralities are present, and *Relative Survival* offers a method to isolate disease-related mortality when cause of death data is unavailable or unreliable. In the context of cancer survival, relative survival often exhibits a plateau, indicating statistical cure. In this paper we use a mixture cure model with Weibull distribution to model relative survival of cancer patients and estimate the proportion of cured cases. We apply our methodology to a cohort of patients diagnosed with one of the most common cancers, colon cancer, and offer insights into their survival. The data is sourced from a long-established Italian population-based cancer registry.

Keywords: cure models · relative survival · survival analysis

1 Introduction

The prognosis of cancer has dramatically improved over the last few decades with an increasing proportion of cancer patients being cured. The proportion of cured of a disease is of interest to patients and clinicians and it is a useful measure to monitor trends in cancer outcome. Furthermore, knowing when a patient can be considered cured from a severe illness enables the re-establishment of the same rights as a person who has never had such condition [1]. Hence, it is of interest to estimate the cure fraction in the diseased population, when also mortality due to other causes competes. Relative Survival (RS) is an easy and established way to isolate mortality due to the disease of interest when cause

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 592–598, 2025.
https://doi.org/10.1007/978-3-031-64431-3_98

of death is unavailable or unreliable. RS is the ratio between the observed (all-cause) survival in a patient group and the survival *expected* in a comparable group in the general population. The expected survival is calculated from the population life tables, after matching for demographic variables such as age, sex, calendar year and possibly some other covariates impacting survival. The underlying assumption is that the disease is *rare*, i.e. that removing the disease of interest from the general population life tables would have negligible effect in the estimated survival. As a result, relative survival provides a measure of *excess mortality* experienced by patients diagnosed with the disease of interest. For some cancers, relative survival hits a plateau after some years, meaning that the mortality rate of a subset of the diseased patients becomes the same as the mortality rate in the general population. When patients no longer experience excess mortality, the group still alive is considered *statistically cured*; this claim only makes sense from a population perspective and should not be mistaken with information on an individual being free of any symptoms related to the disease. Several cure models have been proposed to estimate the proportion of cancer patients that are statistically cured. Within the context of *mixture* cure models, De Angelis et al. [2] employed exponential and Weibull distributions to characterize the survival distribution of fatal individuals, whereas Yu et al. [3] introduced a novel approach by proposing a generalized gamma distribution for the same purpose. Within the context of *non-mixture* cure models, Lambert et al. [4] assumed a parametric Weibull distribution function, whereas Andersson et al. [5] did not make any parametric assumption but rather used cubic splines to model the same distribution function. In this paper we propose an application of the Weibull version of the model from De Angelis et al. on data of colon cancer patients from the Italian population-based cancer registry of Varese, showing how fitting such model per age stratum provides a reasonable fit to the data. Furthermore, differently from the original paper, we use the EM algorithm [6] to find the maximum likelihood estimates of the model parameters. The mixing proportion obtained will provide an estimate of the cured fraction from colon cancer. We compare our parametric models against empirical relative survival ratios estimated using the well-established *Ederer II* method [7], implemented in R [8] via the R-package *relsurv* [9].

2 Methodology

Consider a fatal disease affecting a population and suppose that the individuals can be divided into two distinct subgroups: a proportion P of cured patients with respect to the disease, who have death hazard h_0, density function f_0 and survival probability S_0; and a proportion $(1 - P)$ of patients who are bound to die of the disease of interest, having death hazard h_1, survival probability S_1 and density function f_1, with the following additive assumption for the hazard

$$h_1(t, \boldsymbol{x}) = h_0(t, \boldsymbol{x}) + h_D(t, \boldsymbol{x}), \tag{1}$$

where t and $\boldsymbol{x}$ are the observed failure time and a vector of given covariates respectively, and h_D is the excess risk due to the disease. The corresponding

survival function is given by

$$S_1(t, \boldsymbol{x}) = S_0(t, \boldsymbol{x}) S_D(t, \boldsymbol{x}), \tag{2}$$

where S_D is the survival function associated to h_D, and it is generally referred to as *relative survival ratio* for the disease of interest when S_0 is estimated through population life tables. The use of life tables is justified under the assumption that the mortality due to the *rare* disease is negligible compared to all-cause mortality in the general population.
The cumulative relative survival mixture can be therefore written as

$$S(t, \boldsymbol{x}) = P(\boldsymbol{x}) + (1 - P(\boldsymbol{x})) S_D(t, \boldsymbol{x}), \tag{3}$$

where the relative survival of cured patients is identically equal to 1. The proportion of cured $P(\boldsymbol{x})$ has been modeled as a function of the covariates $\boldsymbol{x}$ using the logistic function:

$$P(\boldsymbol{x}) = 1/(1 + \exp(-\boldsymbol{x}^T \boldsymbol{\pi})), \tag{4}$$

where $\boldsymbol{x}$ is a $(p + 1)$-dimensional vector of covariates with first entry equal to 1 and $\pi \in \mathbb{R}^{p+1}$ are unknown regression parameters. S_D is the relative survival function associated to the uncured group, for which a Weibull parametric form is assumed

$$S_D(t, \boldsymbol{x}) = \exp[-(\lambda(\boldsymbol{x}) t)^{\beta}]. \tag{5}$$

The scale parameter is modeled with a proportional-hazard like model, that is

$$\lambda(\boldsymbol{x}) = \exp(\boldsymbol{x}^T \boldsymbol{\gamma}), \tag{6}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ are again unknown regression parameters. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\gamma}$, and β are estimated by maximum likelihood on the basis of individual data. Let us consider n individuals entering a follow-up of length τ. Let t_i be the observed time to event for the i-Th individual, either censoring time or event time, determined by the censoring indicator δ_i ($\delta_i = 1$ event time and $\delta_i = 0$ censoring time). Then the contribution of the i-Th individual to the likelihood function is given by

$$\begin{aligned} L_i(\boldsymbol{\pi}, \boldsymbol{\gamma}, \beta\,;\, \boldsymbol{x}_i, t_i, \delta_i) = {} & P(\boldsymbol{x}_i)\, S_0(t_i, \boldsymbol{x}_i) h_0(t_i, \boldsymbol{x}_i)^{\delta_i} + \\ & + (1 - P(\boldsymbol{x}_i))\, S_0(t_i, \boldsymbol{x}_i) S_D(t_i, \boldsymbol{x}_i) \cdot \\ & \cdot (h_0(t_i, \boldsymbol{x}_i) + h_D(t_i, \boldsymbol{x}_i))^{\delta_i}; \end{aligned} \tag{7}$$

hence, the log-likelihood for all observations is given by

$$\ell(\boldsymbol{\pi}, \boldsymbol{\gamma}, \beta\,;\, \boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \sum_{i=1}^{n} log(L_i(\boldsymbol{\pi}, \boldsymbol{\gamma}, \beta\,;\, \boldsymbol{x}_i, t_i, \delta_i)). \tag{8}$$

The EM algorithm was used to find the maximum likelihood estimates of the parameters in (8). In particular, to increase the velocity of convergence, the EM algorithm was employed to find suitable starting values of the parameters to be given to the *Optim* function in *R*, performing Quasi-Newton direct maximization of (8), as also suggested in [10]. As a result, the output of the EM was refined and an estimate of the inverse of the observed information matrix obtained.

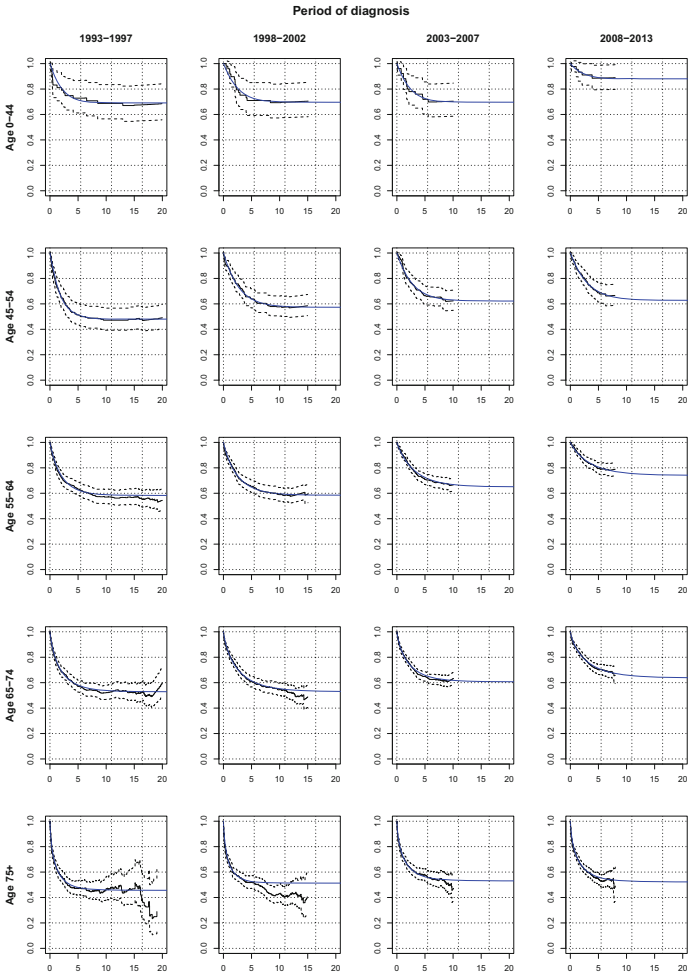


Fig. 1. Relative survival of colon cancer patients by the Ederer II life-time method (black line) with relative 95% confidence interval, and by the estimated mixture models (blue line). Data from 8,647 colon cancer patients diagnosed in Italy from 1993 to 2013. Y-axis represents Relative Survival and the x-axis the follow-up time expressed in years.

3 Experimental Results

The mixture model presented in the previous section was applied to individual colon cancer data collected by the Italian population-based cancer registry of Varese. The cohort consists of all 8,647 patients residing in the municipality of Varese who were diagnosed with colon cancer from 1993 to 2013, followed-up until 31Th December 2013. The data were stratified by age at diagnosis and fitted with age-group specific models considering period of diagnosis as model-specific

covariate. We have considered five age classes: 0–44, 45–54, 55–64, 65–74, 75–99; and four periods: 1993–1997, 1998–2002, 2003–2007, and 2008–2013. Figure 1 shows the relative survival function estimated by age class and period. The black line represents the relative survival rates computed with the Ederer II method with associated confidence intervals (dashed black lines), while the blue line is the estimated mixture model. The model seems to present some issues in fitting the first two periods of the oldest age class, for which the cured assumption is not fully satisfied since no asymptote in the black curve is clearly visible. For all remaining combinations of age and period classes, the model follows very closely the black line, indicating that the assumption of patients being cured of the disease can be confidently made. Table 1 shows, for each combination of age and period of diagnosis, the estimated proportion of cured and the estimated mean survival time for the uncured proportion, with associated standard error. There is a substantial improvement in survival across the examined time periods which is consistent with the diagnostic and therapeutic improvement over the years; the cured proportion is estimated to increase particularly in the second period for the eldest age group. Regarding mean survival time of fatal cases, enhancements in prognosis are evident across all age groups except for the 0–44 and 75–99 categories, where stability persists at approximately 2.5 years for the youngest and 1.5 years for the oldest. This result is consistent with biological differences of tumours diagnosed at young age and with poorer survival due to comorbidities for the elders. Conversely, all other age cohorts show an increase of at least one year, with patients aged 55–64, targeted by screening programs, nearly doubling average time to death. These findings highlight the overall positive trajectory in survival outcomes, with notable variations observed across different age groups.

Table 1. Estimates of proportion of cured and mean survival time of fatal cases (given by $T_m = \lambda^{-1}\Gamma(1 + 1/\beta)$) with corresponding standard errors calculated using delta method ($\varepsilon = 10^{-3}$). Data from 8,647 colon cancer patients in Italy from 1993 to 2013.

Age class	Period of diagnosis			
	1993–1997	1998–2002	2003–2007	2008–2013
Proportion of cured				
0–44	69.1 (7.10)	69.6 (6.91)	69.7 (6.84)	88.0 (5.59)
45–54	48.0 (4.48)	57.3 (4.17)	62.2 (4.23)	62.8 (5.87)
55–64	58.3 (3.05)	58.5 (2.91)	65.0 (2.92)	74.1 (4.38)
65–74	52.8 (2.96)	52.9 (3.09)	60.7 (2.47)	63.7 (3.76)
75+	45.7 (3.35)	51.3 (2.30)	53.1 (3.12)	52.2 (3.07)
Mean survival time for fatal cases (years)				
0–44	2.02 (0.002)	2.41 (0.002)	2.21 (0.002)	2.08 (0.012)
45–54	1.83 ($< \varepsilon$)	2.70 ($< \varepsilon$)	2.66 (0.001)	2.96 (0.001)
55–64	2.12 (0.001)	2.61 ($< \varepsilon$)	3.22 (0.001)	3.64 (0.002)
65–74	2.12 ($< \varepsilon$)	2.90 ($< \varepsilon$)	2.34 (0.001)	3.20 (0.001)
75+	1.23 ($< \varepsilon$)	1.04 (0.001)	1.58 (0.001)	1.69 (0.001)

4 Conclusions

Many cancer patients are actually cured of the disease. Clustering the study population into two different groups - those having the same death risk as the general population, and those experiencing an excess of mortality due to the disease - is therefore a reasonable assumption from a biological point of view. Mixture models allow for simultaneous estimation of the probability of cure and time to death for fatal cases. In this context, we made two key assumptions: first that cure occurs and second that the time to death distribution of the uncured cases can be described by the Weibull distribution. While this assumptions seem to be plausible for young and middle age groups, they are more restrictive for the oldest age group. Hence we could consider applying more flexible approach like the one proposed by Andersson et al. [5] and compare it with the one used in this paper. Furthermore, it would be of interest to consider applying the methodology to various cancer sites across different countries to assess both prognostic differences by cancers and disparities in access to optimal care protocols across different countries.

References

1. Scocca, G., Meunier, F.: Towards an EU legislation on the right to be forgotten to access to financial services for cancer survivors. European J. Cancer (Oxford, England : 1990), **162**, 133–137 (2022)
2. De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., Verdecchia, A.: Mixture models for cancer survival analysis: application to population-based data with covariates. Stat. Med. **18**(4), 441–454. Wiley (1999)
3. Yu, B., Tiwari, R. C., Cronin, K. A., Feuer, E. J.: Cure fraction estimation from the mixture cure models for grouped survival data. Stat. Med. **23**(11), 1733–1747. Wiley (2004)
4. Lambert, P.C., Thompson, J.R., Weston, C.L., Dickman, P.W.: Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics **8**(3), 576–594. Oxford University Press (OUP) (2007)
5. Andersson, T.M., Dickman, P.W., Eloranta, S., Lambert, P.C.: Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC Med. Res. Methodol. **11**(1). Springer Science and Business Media LLC (2011)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Series B (Methodological) **39**(1), 1–38 (1977)
7. Ederer, F., Axtell, L.M., Cutler, S.J.: The relative survival rate: a statistical methodology. Nat. Cancer Inst. Monogr. **6** , 101-121. Bethesda, Maryland, U.S (1961)
8. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (2021)

9. Perme, M.P., Pavlic, K.: Nonparametric relative survival analysis with the R package relsurv. J. Stat. Softw. **87**(8). Foundation for Open Access Statistic (2018)
10. Aitkin, M., Aitkin, I.: A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. Stat. Comput. **6**(2), 127–130. Springer Science and Business Media LLC (1996)

Lower Bias Circular Density Estimation with Contaminated Data

Marco Di Marzio[1(✉)], Stefania Fensore[1], Agnese Panzera[2], and Chiara Passamonti[1]

[1] G. d'Annunzio University, Pescara, PE 65129, Italy
{marco.dimarzio,stefania.fensore,chiara.passamonti}@unich.it
[2] University of Florence, Florence, FI 50134, Italy
agnese.panzera@unifi.it

Abstract. We study the problem of estimating circular densities when sample data are affected by measurement errors. We propose a deconvolution approach involving lower bias kernel estimators which take the additional source of bias due to the presence of measurement errors into account. Some asymptotic properties are discussed, and numerical results are provided.

Keywords: Circular deconvolution · Higher order kernel · Sin-order · von Mises kernel

1 Introduction

Circular data consist of observations defined on the circle with a unit radius. After establishing a zero direction and a sense of rotation, these observations are measured using a periodic scale, typically ranging from $[-\pi, \pi)$ radians or $[0, 360)$ degrees. These data are involved in various scientific fields such as biology, geology, and meteorology. Examples include migration paths, animal directions, wind and marine current directions, as well as orientations of joint faults and oriented stones. For a comprehensive overview of recent advancements in directional statistics, refer to [11,12], and [13].

Data can be affected by errors for various reasons. For instance, errors may stem from the degradation or improper calibration of measurement devices or procedures, as well as challenges in accurately measuring the variables of interest. Furthermore, measurement contamination can occur in various fields including astronomy, meteorology, geology, and econometrics.

For this reason, errors-in-variables context has been extensively studied in the literature, specifically in terms of density and regression estimation both in Euclidean and circular settings (see, for example, [1–4,6]). In this paper, we deepen the circular density estimation problem in the presence of measurement errors by employing some classes of *lower bias circular kernels*.

The paper is organized as follows. In Sect. 2 some basic concepts about kernel circular density estimation in both the error-free and the errors-in-variables

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 599–604, 2025.
https://doi.org/10.1007/978-3-031-64431-3_99

cases, are recalled. Section 3 presents some circular kernels which can achieve lower bias order. Finally, Sect. 4 provides some numerical results on simulated data.

2 Kernel Estimation of Circular Densities

Given a random sample of angles $\Theta_1, \ldots, \Theta_n$ from an unknown circular density f_Θ, the kernel estimator of f_Θ at $\theta \in [0, 2\pi)$ is defined as

$$\hat{f}_\Theta(\theta; \kappa) = \frac{1}{n} \sum_{i=1}^{n} K_\kappa(\Theta_i - \theta), \tag{1}$$

where K_κ is a circular kernel, i.e. a periodic, unimodal, symmetric density function with concentration parameter $\kappa > 0$, which admits a convergent Fourier series representation as follows

$$K_\kappa(\theta) = \frac{1 + 2\sum_{\ell=1}^{\infty} \gamma_\ell(\kappa) \cos(\ell\theta)}{2\pi}.$$

Just as in the classical Euclidean context, the kernel function emphasizes the contribution of the observations located in a neighborhood of the estimation point. The size of this neighborhood is regulated by κ, which acts akin to the inverse of bandwidth in the Euclidean case, meaning that wider neighborhoods correspond to smaller values of κ. Moreover, similarly to the Euclidean setting, the concept of the (sin-) order of a circular kernel is linked to the number of its non-vanishing even (sin-) moments. More specifically, letting $\eta_j(K_\kappa) = \int_0^{2\pi} K_\kappa(u) \sin^j(u) du$, a circular kernel K_κ is said to be a q-th sin-order kernel if

$$\eta_0(K_\kappa) = 1, \quad \eta_j(K_\kappa) = 0 \ \text{ for } j < q, \text{ and } \quad \eta_q(K_\kappa) \neq 0.$$

Popular examples of second sin-order kernels include the *von Mises* density, the *wrapped Normal* and the *wrapped Cauchy* densities.

[5] found that, when K_κ has sin-order q, under suitable smoothness assumptions on f_Θ and assuming that as n increases, for any $\ell \in \mathbb{Z}^+$, $\lim_{n\to\infty} \gamma_\ell(\kappa) = 1$, and $\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{\infty} \gamma_\ell^2(\kappa) = 0$, the resulting kernel estimator satisfies

$$\mathsf{E}[\hat{f}_\Theta(\theta; \kappa)] - f_\Theta(\theta) = \frac{\eta_q(K_\kappa) f_\Theta^{(q)}(\theta)}{q!} + o(1 - \gamma_q(\kappa)), \tag{2}$$

where $f_\Theta^{(q)}(\theta)$ is the derivative of order q of f_Θ evaluated at θ and

$$\mathsf{VAR}[\hat{f}_\Theta(\theta; \kappa)] = \frac{\left(1 + 2\sum_{\ell=1}^{\infty} \gamma_\ell^2(\kappa)\right) f_\Theta(\theta)}{2\pi n} + o\left(\frac{\sum_{\ell=1}^{\infty} \gamma_\ell^2(\kappa)}{n}\right). \tag{3}$$

When $q = 2$, then $\eta_2(K_\kappa) = 2^{-1}(1 - \gamma_2(K_\kappa))$, while for the general case of sin-order $q \geq 2$, [7] showed that $\eta_q(K_\kappa) = O(2^{1-q}(1 - \gamma_q(\kappa)))$.

Now, assume that we wish to estimate the density f_Θ of a random angle Θ, but we have data observed *with errors*, that is, instead of the Θ_is, we have i.i.d. observations $\Phi_1, \ldots, \Phi_n$ with

$$\Phi_i = (\Theta_i + \varepsilon_i) \mathsf{mod}(2\pi),$$

where the ε_is are independent copies of a random angle ε, independent of Θ, whose density f_ε is assumed to be a known circular density symmetric around 0. This is referred to as the classical measurement errors problem, and the density f_Φ of Φ is the circular convolution of f_Θ and f_ε.

The problem of estimating f_Θ then amounts to a circular density deconvolution problem and has been addressed by [3], who defined the following kernel estimator for f_Θ at $\theta \in [0, 2\pi)$

$$\tilde{f}_\Theta(\theta; \kappa) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{K}_\kappa(\Phi_i - \theta), \tag{4}$$

where the weight function

$$\widetilde{K}_\kappa(\phi) := \frac{1}{2\pi} \left\{ 1 + 2 \sum_{\ell=1}^{\infty} \frac{\gamma_\ell(\kappa)}{\lambda_\ell(\kappa_\varepsilon)} \cos(\ell\phi) \right\}$$

is a circular deconvolution kernel, with $\gamma_\ell(\kappa)$ and $\lambda_\ell(\kappa_\epsilon) \neq 0$ (for each ℓ) respectively being the ℓ-th Fourier coefficient of a circular kernel K_κ, and the ℓ-th Fourier coefficient of the error density f_ε (with concentration κ_ε). An interesting property of estimator (4) is that, as it happens in the Euclidean setting, it shares the asymptotic bias of estimator (1). Then, using a circular kernel of sin-order q for K_κ, the deconvolution kernel $\widetilde{K}_\kappa$ has sin-order q, and the resulting estimator attains the asymptotic bias in (2). Concerning the asymptotic variance of estimator (4), it depends on the degree of smoothness of the error term, corresponding to the asymptotic variance in (3) but with $\gamma_\ell(\kappa)$ replaced by the ratio $\gamma_\ell(\kappa)/\lambda_\ell(\kappa_\varepsilon)$, $\ell \in \mathbb{Z}^+$.

3 Small Biased Von Mises-Based Circular Kernels

The von Mises kernel is defined by

$$K_\kappa(\theta) = \frac{e^{\kappa \cos(\theta)}}{2\pi \mathcal{I}_0(\kappa)}, \quad \theta \in [0, 2\pi)$$

where $\mathcal{I}_\ell(k)$, $\ell \in \mathbb{Z}$, is the modified Bessel function of order ℓ, and its ℓ-th Fourier coefficient is $\gamma_\ell(k) = \mathcal{I}_\ell(k)/\mathcal{I}_0(k)$. As a result, when K_κ is a von Mises kernel, in both the error-free and the errors-in-variables cases, the resulting kernel density estimators have a leading bias $4^{-1}(1 - \mathcal{I}_2(\kappa)/\mathcal{I}_0(\kappa)) f_\Theta^{(2)}$. Now, for a big enough κ it holds that $2^{-1}(1 - \mathcal{I}_2(\kappa)/\mathcal{I}_0(\kappa)) \sim 1/\kappa$, and then the bias of the resulting kernel density estimators is $O(1/\kappa)$.

[7] observed that higher sin-order kernels do not necessarily lead to a reduced bias, unlike in the Euclidean context. Conversely, certain classes of kernels, which are derived from the von Mises one, yield a lower bias even with a second sin-order.

Higher sin-order circular kernels, which have been also studied in a more general setting by [6,7], can be obtained starting from second-order ones by adapting the building scheme of higher order Euclidean kernel in [10]. This leads to the following Lejeune and Sarda (LS) type circular kernel

$$K_\kappa^{LS}(\theta) := \frac{\left\{\eta_4(K_\kappa) - \sin(\theta)^2\eta_2(K_\kappa)\right\} K_\kappa(\theta)}{\eta_4(K_\kappa) - \eta_2^2(K_\kappa)}.$$

The kernel $K_\kappa^{LS}(\theta)$ is a fourth sin-order kernel and when K_κ is a von Mises, it holds that, for a big enough κ, $\eta_4(K_\kappa^{LS}) \sim -1/\kappa^2$, so the asymptotic bias of the resulting kernel density estimator is $O(1/\kappa^2)$.

Further types of circular kernels, respectively based on the Euclidean counterparts in [8,9,14], and also studied in a more general setting by [6], are

- the twicing (TW) type circular kernel

$$K_\kappa^{TW}(\theta) := 2K_\kappa(\theta) - K_{\kappa/2}(\theta),$$

- the Hall and Marron (HM) type circular kernel is

$$K_\kappa^{HM}(\theta) := \frac{1}{2}\left(3K_\kappa(\theta) + \sin(\theta)K_\kappa^{(1)}(\theta)\right),$$

- the Jones and Foster (JF) type circular kernel

$$K_\kappa^{JF}(\theta) := K_\kappa(\theta) - \frac{1}{2}\eta_2(K_\kappa)K_\kappa^{(2)}(\theta).$$

When K_κ is a von Mises kernel, all these kernels, even with a second sin-order, produce a $O(1/\kappa)$ bias.

Denoting as $\gamma_\ell^s(\kappa)$, $s \in \{LS, TW, HM, JF\}$, the ℓ-th Fourier coefficient of the kernel K_κ^s, the deconvolution version $\widetilde{K}_\kappa$ is obtained by using the ratio $\gamma_\ell^s(\kappa)/\lambda_\ell(\kappa_\varepsilon)$ in its Fourier series representation. Obviously, the use of all these kernels produces a $O(1/\kappa^2)$ bias of the resulting kernel estimator for the errors-in-variables context.

4 Simulations

In our simulation study, we consider two simple models where the target population f_Θ is the von Mises density (vM), and the error distribution f_ε is modeled as the wrapped Laplace (wL). We hypothesize two levels of measurement errors, according to the value of the noise-to-signal ratio (NSR), i.e. the ratio between the *circular variance* of ε and that of Θ. Specifically, the considered scenarios, which are depicted in Fig. 1, are the following:

– Scenario 1: target density vM$(\pi, 1)$, error density wL$(0, .40)$, NSR = 25%
– Scenario 2: target density vM$(\pi, 1.3)$, error density wL$(0, .50)$, NSR = 45%

Notice that, other than the locations, for the von Mises density we have a concentration parameter, while for the wrapped Laplace one, we have a spread measure.

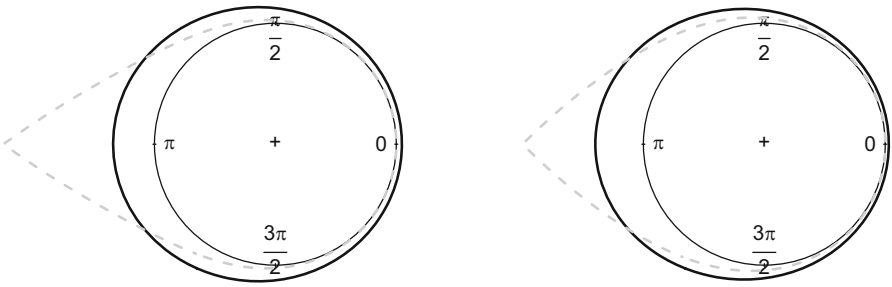


Fig. 1. Scenarios 1 (left) and 2 (right): target population (continuous black) and error density (dashed gray). The common densities support is represented by the innermost circle. For illustrative purposes, both the target and error models are centered on π.

We draw from the above models 500 samples of size $n = 100$ and $n = 1000$, and compare the deconvolution estimator employing the von Mises kernel (DEC) and that one involving the lower bias kernels (TW, JF, LS, HM), with respect to the naive kernel density one (KDE), when K_κ is von Mises. Deconvolution estimates do not give bona-fide densities for involving small bias kernels, therefore they have been clipped and normalized.

Our performance index is the averaged integrated squared error (AISE). In order to isolate specific merits of the estimators, in this study we use for all the methods the best possible smoothing degree obtained through a grid search. The results are collected in Table 1. We notice that the deconvolution estimators generally improve on the naive approach. Quite interestingly, the proposed methods improve even on the deconvolution estimator, still giving similar performances for small sample sizes. This appears very encouraging if considered

Table 1. AISE (x1000) over 500 samples of size n in Scenario 1 (top panel) and Scenario 2 (bottom panel).

n	KDE	DEC	TW	JF	LS	HM
100	1.42	1.38	1.12	1.43	1.48	1.33
1000	0.55	0.47	0.41	0.43	0.41	0.43
100	2.38	2.18	1.89	2.20	2.11	2.13
1000	1.32	1.20	1.10	1.11	1.10	1.12

that the traditional theory of higher order kernels affirms that improvements are possible only for big sample sizes. Surely, additional simulative scenarios need to be considered in order to confirm such evidence.

References

1. Delaigle, A.: An alternative view of the deconvolution problem. Stat. Sin. 1025-1045 (2008)
2. Delaigle, A.: Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes. Aust. N. Z. J. Stat. **56**(2), 105–124 (2014)
3. Di Marzio, M., Fensore, S., Panzera, A., Taylor, C.C.: Density estimation for circular data observed with errors. Biometrics **78**(1), 248–260 (2022)
4. Di Marzio, M., Fensore, S., Taylor, C. C.: Kernel regression for errors-in-variables problems in the circular domain. Stat. Methods Appl. 1-21 (2023)
5. Di Marzio, M., Panzera, A., Taylor, C.C.: Local polynomial regression for circular predictors. Stat. Probab. Lett. **79**(19), 2066–2075 (2009)
6. Di Marzio, M., Fensore, S., Panzera, A., Passamonti, C.: On estimating the difference between two circular densities in presence of measurement error. Submitted (2024)
7. Di Marzio, M., Panzera, A., Taylor, C.C.: Kernel density estimation on the torus. J. Stat. Plan. Infer. **141**(6), 2156–2173 (2011)
8. Hall, P., Marron, J.S.: Variable window width kernel estimates of probability densities. Probab. Theory Relat. Fields **80**(1), 37–49 (1988). https://doi.org/10.1007/BF00348751
9. Jones, M.C., Foster, P.J.: Generalized jackknifing and higher order kernels. J. Nonparametric Stat. **3**(1), 81–94 (1993)
10. Lejeune, M., Sarda, P.: Smooth estimators of distribution and density functions. Comput. Stat. Data Anal. **14**(4), 457–471 (1992)
11. Ley, C., Verdebout, T.: Modern Directional Statistics. CRC Press, Boca Raton (2017)
12. Ley, C., Verdebout, T.: Applied Directional Statistics: Modern Methods and Case Studies. CRC Press, Boca Raton (2018)
13. Mardia, K.V., Jupp, P.E.: Directional Statistics. John Wiley & Sons Press, Chichester (2009)
14. Stuetzle, W., Mittal, Y.: Some comments on the asymptotic behavior of robust smoothers. In: Smoothing Techniques for Curve Estimation. Proceedings, Heidelberg 1979, Lecture Notes in Mathematics, vol.757. Springer-Verlag, Berlin, 191–195 (1979)

A New Methodological Approach to Build a Multidimensional Inequality Composite Indicator

Cinzia Di Nuzzo(✉), Venera Tomaselli, and Gianpiero Torrisi

Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania, Italy
cinzia.dinuzzo@unict.it

Abstract. This study introduces a methodological framework for creating a MultiDimensional Inequality Gini Composite Indicator (MDIGI) to assess cultural and digitalization disparities among Italian regions. The novel approach involves extending PCA to three-way data using the PARAFAC model to determine indicator weights, thereby enhancing the MDIGI's capacity to capture inequalities across multiple dimensions in composite indicator construction.

Keywords: PARAFAC model · Multidimensional Inequality Gini indicator

1 Introduction

Composite indicators play a crucial role in the assessment of multifaceted concepts that defy singular measurements. Indicators facilitate comparisons more efficiently than simple differences would allow and are particularly crucial when analysing phenomena with temporal variations. Additionally, they are commonly used to describe territorial series, comparing the values assumed by variables in different territorial areas. Composite indicators serve as tools for comparing multiple statistical units concerning complex socio-economic phenomena, quantified through an appropriate aggregation of elementary indicators.

The outcome of a composite indicator is the ability to rank statistical units based on the measures assigned to the complex phenomenon. In practice, constructing a composite indicator aims to compare different statistical units through the quantification, via a multidimensional phenomenon and analysis, of aspects that need to be grasped simultaneously. Therefore, a composite indicator is the result of aggregating elementary indicators.

One of the main issues in building a composite indicator is to find a way to weigh the statistical units. Specifically, in this work, an objective approach has been considered as it confers several advantages: firstly, grounded in clear and measurable criteria, enhances the transparency and comprehensibility of the weight assignment process, facilitates the reproducibility of analyses by other

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 605–610, 2025.
https://doi.org/10.1007/978-3-031-64431-3_100

researchers or analysts, this contributes to ensuring consistency in evaluations; secondly, an objective approach minimises the risk of introducing bias into weight assignments. Subjective approaches may be susceptible to personal opinions or expert preferences, whereas an objective approach relies on empirical evidence or measurable criteria, thereby mitigating subjectivity.

Specifically, the use of Principal Component Analysis (PCA) to weigh these indicators has garnered significant attention within various domains, with proponents advocating for its utility in aggregating diverse variables into comprehensive measures. PCA has been extensively employed to develop composite indicators across diverse sectors. The intrinsic capability of PCA as a variable reduction technique, especially in highly correlated scenarios, underscores its suitability in condensing numerous observed variables into a compact set of principal components, effectively capturing the variance in the data.

In this framework, our goal is to build a multidimensional inequality Gini indicator that is paramount in comprehensively assessing and addressing societal disparities across diverse dimensions. The conventional Gini coefficient, primarily applied to quantify income inequality, offers a unidimensional perspective, often overlooking variations in different dimensions. By extending the Gini framework to multiple dimensions, a more nuanced and holistic understanding of inequality emerges, enabling policymakers and researchers to capture the multifaceted nature of disparities within a population, see [2]. Therefore, to build a multidimensional inequality Gini Index, we propose addressing methodological challenges by integrating weighting procedures using an extension of PCA for three-way data, following the multidimensional Inequality Gini index's PARAFAC model as proposed by [1,3].

The rest is organized as follows: in Sect. 2 the methodology is formalized and in Sect. 3 a brief sketch of a real application of social sustainability framework regarding the disparities among Italian regions on culture participation and digitalization is provided.

2 Multidimensional Inequality Gini Indicator

In our proposal, we have examined the multidimensional Gini inequality indicator as defined by [2]. Specifically, it is formulated according to the following expression

$$I_G(X) = 1 - \frac{\sum_{i=1}^{N} \left[\left(\frac{r^i}{N} \right)^{\delta} - \left(\frac{r^i - 1}{N} \right)^{\delta} \right] \left(\sum_{j=1}^{M} w_j \left(x_{ij} \right)^{\beta} \right)^{1/\beta}}{\left(\sum_{j=1}^{M} w_j \mu \left(x_j \right)^{\beta} \right)^{1/\beta}} \tag{1}$$

where $X = (x_{ij})$, for $i = 1, \ldots, N$ and $j = 1, \ldots, M$, is the data matrix consists of N objects and M variables; δ and β are two positive parameters; w_j are weights such that $w_j > 0$ and $\sum_{j=1}^{M} w_j = 1$; r^i is a shorthand for $r^i(s)$, the rank of object i on the basis of s, with $s^i = \left(\sum_{j=1}^{M} w_j \left(x_{ij} \right)^{\beta} \right)^{1/\beta}$; and $\mu(x_j)$ is the

mean of the column x_j. Moreover, according to the reference literature [2] two optimal choices for the parameters are $\beta < 1$ and $1 \leq \delta \leq 2$.

The goal of our proposal is to build a MultiDimensional Inequality Gini Composite Indicator (MDIGI) that can provide a measure of inequality among statistical units given a certain time interval. Therefore, the methodology used here aims to build a multidimensional inequality Gini indicator, which takes into account the information of variables considered for each statistical unit (for example the Italian regions) and each occasion (for example for each year). This is important to provide a measure that summarizes the inequalities over the entire period being considered. In our proposal, the weights are chosen by using the PARAFAC model [1,3] that can be viewed as an extension of PCA to three-way data. Indeed, frequently, in real applications, we confront data with three dimensions, so adaptations of PCA tailored for such three-way data are recognized as three-way PCA. This variant has garnered significant attention due to its relevance and applicability in handling multi-dimensional datasets.

2.1 A New Methodology to Build a Multidimensional Inequality Gini Indicator

The main contribution of this work lies in formulating a novel multidimensional Gini inequality indicator that incorporates multiple dimensions of the data, estimating the weights via the PARAFAC model, see [1,3].
A three-way data array can be denoted as

$$\boldsymbol{X} = \{x_{ijh},, \quad for \quad i = 1, \ldots, N, \quad , j = 1, \ldots, M, \quad h = 1, \ldots H\}. \tag{2}$$

For example, x_{ijh} could stand for the (i, j) element recorded at time h. A popular procedure is PARAFAC, which was proposed by [1,3]. The PARAFAC model used here (i.e. with rank equal to 1) can be described by

$$\boldsymbol{X}_h = \boldsymbol{A}\boldsymbol{C}_h\boldsymbol{B}^\top + \boldsymbol{E}_h, \qquad for \quad h = 1, \ldots, H. \tag{3}$$

where $\boldsymbol{A}$ is a N-vector that represents the scores of objects; $\boldsymbol{B}$ is a M-vector that represents the scores of the variables; and $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_H$ represents the scores of the occasions.

In order to extend (1) and to provide the weights of the multidimensionality inequality Gini Indicator, we consider the following products

$$\overline{\boldsymbol{W}} = \boldsymbol{B}\boldsymbol{C}^\top. \tag{4}$$

Since the weights of the Gini indicator have to sum to one, we normalize the previous quantity as follows

$$w_{jh} = \frac{\overline{w}_{jh}}{\sum_{m=1}^{M} \overline{w}_{mh}}, \quad for \quad j = 1, \ldots, M, \quad h = 1, \ldots, H. \tag{5}$$

Therefore, one can build an $M \times H$ matrix of the weights as

$$\boldsymbol{W} = (w_{jh}), \quad for \quad j = 1, \ldots, M, \quad h = 1, \ldots, H. \tag{6}$$

So, following the same ideas of [2], the scores of the multidimensional inequality Gini indicator are

$$s_{ih} = \left(\sum_{m=1}^{M} w_{jh} \left(x_{ijh} \right)^{\beta} \right)^{1/\beta}. \tag{7}$$

Afterwards, we sort s_i in non-increasing order for all $h = 1, \ldots, H$.

Let r_{ih} be the rank of object i at the occasion h based on s_{ih}. Let $R = (r_{ih})$ be the $N \times H$ matrix of the ranks for $i = 1, \ldots, N$ and $h = 1, \ldots, H$. Let $\boldsymbol{\mu} = (\mu_{jh})$ be the means matrix, where $\mu_{jh} = \frac{\sum_{i=1}^{N} x_{ijh}}{N}$.

The multidimensional inequality Gini Indicator for three-way data (MDIGI3W) is defined as

$$I_h(\boldsymbol{X}) = 1 - \frac{\sum_{i=1}^{N} \left[\left(\frac{r_{ih}}{N} \right)^{\delta} - \left(\frac{r_{ih}-1}{N} \right)^{\delta} \right] \left(\sum_{j=1}^{M} w_{jh} \left(x_{ijh} \right)^{\beta} \right)^{1/\beta}}{\left(\sum_{j=1}^{M} w_{jh} \mu_{jh}^{\beta} \right)^{1/\beta}}, \quad for \quad h = 1, \ldots, H, \tag{8}$$

where δ and β have the same role of (1).

The multidimensional inequality Gini indicator defined in (8) gives a measure of the disparities among statistical objects for each occasion, this is useful to compare how to change the disparities of socio-economic variables in a certain time interval. However, it can be important to define an MDIGI that provides a unique measure of the disparities among objects, variables, and occasions simultaneously. Therefore, similarly, following the same line of thought of (8), we have defined a second definition of the *Total MDIGI3W* is

$$I(\mathbf{X}) = 1 - \frac{\sum_{i=1}^{N} \left\{ \left[\left(\frac{r_i}{N} \right)^{\delta} - \left(\frac{r_i - 1}{N} \right)^{\delta} \right] \left(\sum_{j=1}^{M} w_j \left(\tilde{x}_{ij} \right)^{\beta} \right)^{\frac{1}{\beta}} \right\}}{\left(\sum_{j=1}^{M} w_j \mu(\tilde{\mathbf{x}}_j)^{\beta} \right)^{\frac{1}{\beta}}}, \tag{9}$$

where according to the PARAFAC approach (3), $w_j = b_j$ are weights such that $w_j > 0$ and normalized to have $\sum_{j=1}^{M} w_j = 1$; $\tilde{\mathbf{X}} = \mathbf{X}_i \mathbf{C}$ is a matrix of size $N \times M$, where $\mathbf{X}_i$ is the slice of three-way data of size $M \times H$ and $\mathbf{C}_1, \ldots, \mathbf{C}_H$ are the H-vectors of the scores on the occasions; r_i is a shorthand for $r_i(s)$ the ranks of unit i on the basis of $s^i = \left(\sum_{j=1}^{M} w_j \left(\tilde{x}_{ij} \right)^{\beta} \right)^{\frac{1}{\beta}}$; and, $\mu(\tilde{\mathbf{x}}_j)$ is the mean of the column $\tilde{\mathbf{x}}_j$.

3 An Application on Cultural Participation and Digitalization Across Italian Regions

This study aims to employ the statistical methodology outlined in Sect. 2 for data processing and aggregation, to construct the multidimensional inequality Gini composite indicator to assess territorial dynamics regarding cultural participation and digitalization at the regional level in Italy. This application serves as an

initial assessment of disparities in Social Sustainability among Italian regions. Consumption and cultural participation are integral to social cohesion, with emerging technologies reshaping traditional modes of cultural engagement. Consequently, cultural participation and digitalization represent crucial domains in the analysis of Social Sustainability. To construct a composite indicator for Cultural Participation and Digitalization spanning the years 2011 to 2022 in Italian regions, particular attention has been given to variables composed of 10 variables for *digitalization* that are regular internet usage, internet households, internet users, internet-access households, internet usage in the last 3 months, broadband connections, daily internet use, internet use at least once a week in the last 3 months, internet banking activity in the last 3 months, and online shopping in the last 3 months; and, 10 variables for *culture* that are cultural participation outside the home, consumption of books and newspapers, weekly newspaper readership, daily newspaper readership, annual book consumption, book consumption ranging from 1 to 3 books per year, annual cinema attendance, annual museum and exhibition visits, annual participation in sports activities, and annual visits to archaeological sites. According to the literature, two optimal choices for the parameters are $\beta = 0.1$ and $\delta = 1$.

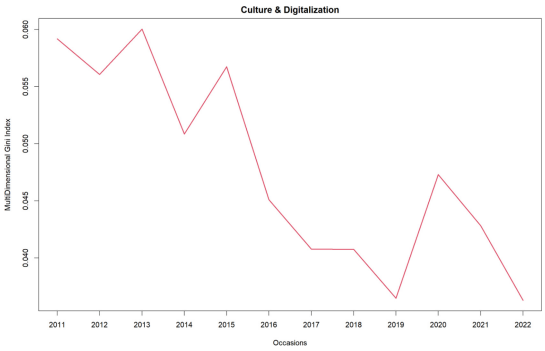


Fig. 1. Plot of MDIGI3W 2011–2022.

In Fig. 1, the values of the MDIGI3W Gini indicator, constructed according to the formula in (8), are displayed over time. As observed from the graph, the indicator values exhibit a declining trend while remaining relatively similar. Thus, it can be asserted that the inequalities among Italian regions regarding the macro area of cultural participation and digitalization are gradually decreasing. Within this decreasing trend, however, certain local maximum peaks are noticeable in the years 2013, 2015, and 2020. It is evident that inequalities among Italian regions increased in 2020 due to the COVID-19 pandemic, and for instance, in 2015, regional disparities in terms of culture were influenced by the presence of EXPO in Lombardia.

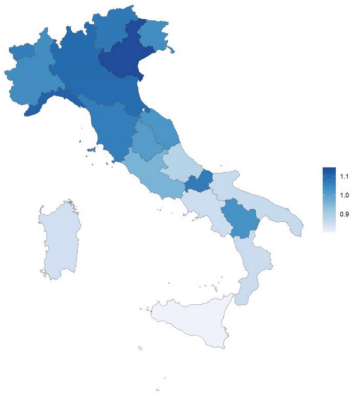


Fig. 2. Scores of *Total MDIGI3W* 2011–2022.

To conduct further analysis and comprehend the ranking of Italian regions in constructing the indicator, Fig. 2 illustrates the scores of the total MDIGI3W indicator over the period 2011–2022. From the figure, a clear configuration emerges wherein Italian regions are grouped into northern, central, and southern regions. It is discernible that northern regions generally hold more weight in assessing the Gini indicator compared to southern regions.

Acknowledgement. This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018—CUP E63C22002120006).

References

1. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika **35**(3), 283–319 (1970)
2. Decancq, K., Lugo, M. A.: Inequality of wellbeing: a multidimensional approach. Economica **79**(316), 721 746 (2012)
3. Harshman, R. A.: Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi mode factor analysis. UCLA Working Pap. Phonetics **16**(1) 84 (1970)



Regularized Functional Partial Least Squares Regression of Neuroimaging Data

Pietro Donelli[1,2(✉)], Harold A. Hernández-Roig[4,5], M. Carmen Aguilera-Morillo[6], Eleonora Arnone[3], Letizia Clementi[7], Rosa E. Lillo[4,5], Alessandro Palummo[1], and Laura M. Sangalli[1]

[1] Politecnico di Milano, DMEC - Department of Mechanics, Milano, Italy
pietro.donelli@polimi.it
[2] Politecnico di Milano, MOX - Department of Mathematics, Milano, Italy
[3] Department of Management, Università degli Studi di Torino, Turin, Italy
[4] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
[5] uc3m-Santander Big Data Institute, Madrid, Spain
[6] Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Valencia, Spain
[7] Fondazione IRCCS Istituto Neurologico Carlo Besta, SC Neurologia, Salute Pubblica, Disabilità, Milan, Italy

Abstract. Thanks to the rapid technological advancement across scientific and engineering domains, the acquisition of extensive and complex data has become increasingly feasible. One notable source of such data is Functional Magnetic Resonance Imaging (fMRI), which allows for the observation of brain activity in a dynamic and spatial context. The analysis of fMRI data, which are defined over the complicated geometry of the brain, poses unique challenges and opportunities. The advent of fMRI has significantly impacted neuroscience, allowing researchers to pinpoint brain regions involved in specific cognitive activities. However, the existing statistical methods for Functional Data Analysis are often confined to Euclidean domains. This paper introduces an innovative approach to functional Partial Least Squares regression designed for scalar-on-function regression problems. The use of a differential regularization term to accommodate the functional nature of data makes the proposed model particularly suitable for handling data defined over complex domains.

Keywords: Functional Data Analysis · Nonparametric Statistics · Differential Regularization · Non-Euclidean Domains · Functional Connectivity Maps

1 Introduction

This work addresses the functional regression problem for data observed over a complex domain $\mathcal{D}$, utilizing functional Partial Least Squares (fPLS) regression. Specifically, this paper focuses on the scalar-on-function regression problem,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 611–616, 2025.
https://doi.org/10.1007/978-3-031-64431-3_102

where the predictor variable is a random field $\mathcal{X} : \mathcal{D} \rightarrow \mathbb{R}$, and the response is a (possibly multivariate) random variable $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^L$. To tackle the scalar-on-function regression problem, [13] introduced the fPLS approach as a generalization of Partial Least Squares (PLS) regression to the scenario of functional data. For an introduction to fPLS and functional Principal Components Analysis-based regression, see [6].

The use of a basis representation is common in the Functional Data Analysis literature due to the infinite-dimensional nature of functional data. Common approaches in this field involve presmoothing steps, where the infinite-dimensional variables are projected onto a finite-dimensional functional space spanned by basis functions such as B-splines. In the context of scalar-on-function fPLS regression, various solutions combining B-spline basis expansions and roughness penalties exist in the literature [1,2,10,14]. Such methods focus on functional data defined over Euclidean domains, typically intervals in $\mathbb{R}$.

Here we present an innovative fPLS method, named fPLS-PDE, that draws inspiration from the regularized functional Principal Components Analysis (fPCA) approaches introduced by [3,4,9,11,12]. In contrast with existing fPLS methods, the proposed approach operates directly on the noisy and discrete functional data observations without performing any presmoothing step. Moreover, it accommodates the analysis of functional data defined over multidimensional non-Euclidean domains, such as two-dimensional (2D) Riemannian manifolds and non-convex volumes. The method employs a differential regularization term to penalize the roughness of functional estimates. This paradigm shift ensures smoothness where required, i.e. in the functional estimates, eliminating the need for preprocessing the data and avoiding the potential loss of possibly important information at the early stages of the analysis.

We are motivated by a real-world data application that aims to discriminate patients with schizophrenia from healthy individuals by analyzing their brain connectivity maps. In this application, the connectivity maps are considered as functional data defined over the gray matter: a three-dimensional (3D) non-convex domain $\mathcal{D}$. By employing a Finite Element discretization of the 3D space, the proposed fPLS-PDE method can model the data while accounting for the complex structure of $\mathcal{D}$. Figure 1 shows two views of the resulting FE triangulation of the template brain considered. The upper view and mid-sagittal cut reflect its folded structure and internal cavities.

2 Functional Partial Least Squares Regression

Partial Least Squares encompasses a spectrum of techniques that establish connections between observed variables through latent variables. At the core of all PLS techniques is the fundamental premise that observed data result from a system or process influenced by a limited number of latent variables that cannot be directly measured or observed.

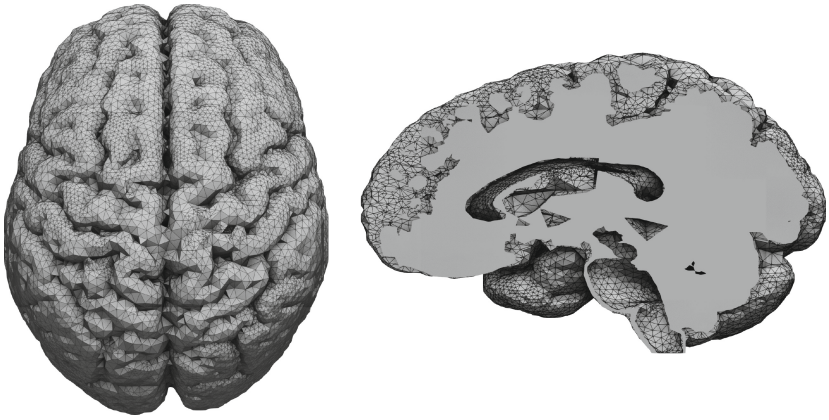


Fig. 1. Left: upper view of the triangulation of a 3D brain using a mesh with 36035 nodes. Right: a mid-sagittal cut (the slice has been shaded in gray) revealing the internal cavities in the right hemisphere.

PLS regression aims to distill predictors into a set of latent components (scores) to reduce the dimensionality of the regression problem. Subsequently, the model utilizes these estimated latent scores for regression. In fact, the term "Partial" in Partial Least Squares underlines that the model extracts only a subset of information from predictors to model the response variable. For an in-depth understanding of PLS regression, refer to [8,15,16].

Let $x_1, \ldots, x_N$ be N noisy realization of a random field $\mathcal{X}(\boldsymbol{p})$ observed at a common set of discrete observations points $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_S\} \in \mathcal{D}$. Let X be the $N \times S$ matrix of predictors with entries $X_{ij} = x_i(\boldsymbol{p}_j)$. L scalar responses associated with each realization of the field organized in a $N \times L$ matrix Y are also given.

Although fPLS is an extension of PLS regression, its structure remains the same as its multivariate counterpart. In particular, it can be divided into three steps repeated sequentially to estimate the subsequent components:

1. **Covariance maximization**: the latent component scores $\boldsymbol{t} \in \mathbb{R}^N$ are identified as the linear combinations of the original variables that capture the maximum covariance between X and Y.
2. **Regression**: the computed scores are used as a new predictor variable to build two regression models for explaining Y and X, respectively.
3. **Deflation**: the estimates obtained in the second step are subtracted from the original variables.

In the proposed fPLS-PDE method, the functional nature of data is accounted for by a differential regularization term added to the loss functions in steps 1 and 2, which involve the estimation of functional unknowns. Unlike what is usually done in the fPLS literature, such a differential regularization term directly fosters smoothness of the functional estimates of interest without the need for any presmoothing step.

3 Application to Neuroimaging

In this section, we briefly comment upon the application of the fPLS-PDE model to neuroimaging data obtained from a task-based fMRI experiment involving healthy individuals and patients with schizophrenia. The dataset includes functional connectivity maps obtained by processing fMRI scans from the Consortium for Neuropsychiatric Phenomics at UCLA. In this process [7], the middle frontal gyrus of the left hemisphere is used as a Region of Interest (ROI). The intricate anatomy of the brain, characterized by gyri and sulci, makes traditional analysis challenging. On the other hand, the proposed fPLS-PDE method provides an ideal framework for functional data observed in such a complex 3D domain, effectively aligning with its geometry.

The application reveals that the latent space obtained from fPLS-PDE allows for perfect classification of healthy and schizophrenia groups, indicating its effectiveness. Three components are sufficient for clear separation, and data reconstruction enables the generation of characteristic connectivity maps for each group. Notably, patients with schizophrenia exhibit a pronounced hypo-frontality in these connectivity maps, aligning with existing literature [5] and highlighting the potential of fPLS-PDE for analyzing complex neuroimaging data (Fig. 2).

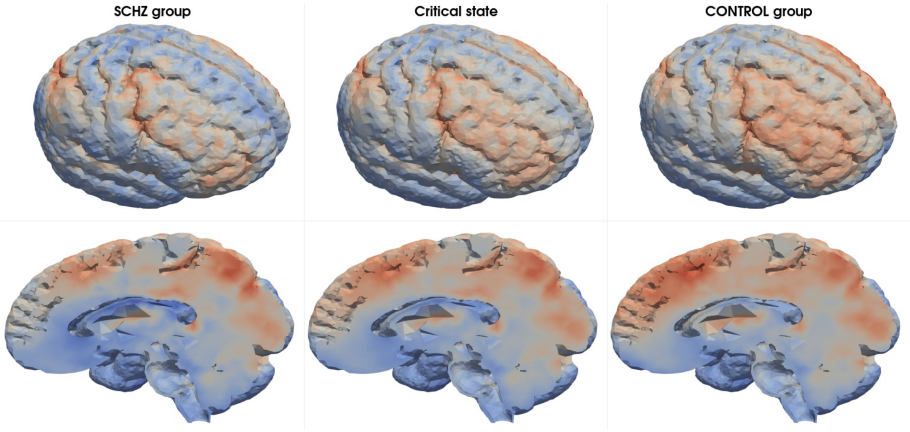


Fig. 2. Characteristic connectivity maps of the considered groups.

4 Conclusions

In this work, we introduced a novel approach to functional Partial Least Squares regression, which incorporates a penalized differential regularization term to account for the functional nature of data. This method is specifically designed to handle scalar-on-function regression problems over complex domains, making it well-suited for applications in neuroimaging. The regularization term, inspired by regularized functional Principal Components Analysis, allows for the analysis

of functional data defined over multidimensional non-Euclidean domains, such as 2D Riemannian manifolds and non-convex volumes.

The application of fPLS-PDE to neuroimaging data showcased the method's efficacy, for instance, in discriminating patients with schizophrenia from healthy individuals using brain connectivity maps. The latent space obtained through fPLS-PDE reveals distinct connectivity patterns, providing physiological insights that corroborate findings in the literature.

Overall, the presented fPLS-PDE methodology opens new avenues for analyzing functional data in diverse fields, particularly in neuroscience and medicine, where understanding complex relationships between functional data is crucial.

Acknowledgments. A. Palummo acknowledges funding by European Union - NextGenerationEU. We also acknowledge the MUR research project Dipartimento di Eccellenza 2023–2027, Dipartimento di Matematica, Politecnico di Milano.

References

1. Aguilera, A., Aguilera-Morillo, M., Preda, C.: Penalized versions of functional PLS regression. Chemom. Intell. Lab. Syst. **154**, 80–92 (2016)
2. Aguilera, A.M., Escabias, M., Preda, C., Saporta, G.: Using basis expansions for estimating functional PLS regression. Chemom. Intell. Lab. Syst. **104**(2), 289–305 (2010)
3. Arnone, E., Negri, L., Panzica, F., Sangalli, L.M.: Analyzing data in complicated 3D domains: smoothing, semiparametric regression, and functional principal component analysis. Biometrics (2023). https://doi.org/10.1111/biom.13845
4. Clementi, L., Arnone, E., Santambrogio, M.D., Franceschetti, S., Panzica, F., Sangalli, L.M.: Anatomically compliant modes of variations: new tools for brain connectivity. PLOS ONE **18**(11), e0292450 (2023)
5. Dove, A., Pollmann, S., Schubert, T., Wiggins, C.J., Yves Von Cramon, D.: Prefrontal cortex activation in task switching: an event-related fMRI study. Cogn. Brain Res. **9**(1), 103–109 (2000)
6. Febrero-Bande, M., Galeano, P., González-Manteiga, W.: Functional principal component regression and functional partial least-squares regression: an overview and a comparative study. Int. Stat. Rev. **85**(1), 61–83 (2017)
7. Gorgolewski, K.J., Durnez, J., Poldrack, R.A.: Preprocessed consortium for neuropsychiatric phenomics dataset. F1000Res. **6**, 1262 (2017)
8. Höskuldsson, A.: PLS regression methods. J. Chemom. **2**(3), 211–228 (1988)
9. Huang, J.Z., Shen, H., Buja, A.: Functional principal components analysis via penalized rank one approximation. Electron. J. Stat. (2008). https://doi.org/10.1214/08-EJS218
10. Krämer, N., Boulesteix, A.L., Tutz, G.: Penalized partial least squares with applications to B-spline transformations and functional data. Chemom. Intell. Lab. Syst. **94**(1), 60–69 (2008)
11. 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. (2016). https://doi.org/10.1214/16-aoas975
12. 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

13. Preda, C., Saporta, G.: Clusterwise PLS regression on a stochastic process. Comput. Stat. Data Anal. **49**(1), 99–108 (2005)
14. Reiss, P.T., Ogden, R.T.: Functional principal component regression and functional partial least squares. J. Am. Stat. Assoc. **102**(479) (2007)
15. Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J., (eds.) Subspace, Latent Structure and Feature Selection, vol. 3940, pp. 34–51. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
16. Tenenhaus, M., Gauchi, J.P., Ménardo, C.: Régression PLS et applications. Revue de statistique appliquée XLIII **1**, 7–63 (1995)

Non-intact Families and Children's Vaccination Coverage: Evidence from Italy

Maria Veronica Dorgali[1](✉), Raffaele Guetto[1], Valentina Tocchioni[1], Alice Dominici[2], and Daniele Vignoli[1]

[1] University of Florence, P.Za Di San Marco, 4, 50121 Florence, Italy
{mariaveronica.dorgali,raffaele.guetto,valentina.tocchioni,daniele.vignoli}@unifi.it
[2] Bocconi University, Via Roberto Sarfatti, 25, 20100 Milan, Italy
alice.dominici@unibocconi.it

Abstract. In recent decades, there has been a notable rise in separations and divorces in many Western countries, leading to an increase in children growing up in non-intact families. Despite the extensive literature on social and economic consequences for children living in non-impact families, a noticeable gap exists in understanding how crucial events like separation influence parents' attitudes towards children's vaccination schedules.

This work analyses whether family arrangements affect children's pediatric vaccination in Italy, testing if parents heading non-intact families have a lower propensity to vaccinate their children than those with intact families. Our sample includes 8,455 parents aged 40 and above residing in Italy, surveyed in 2023. Preliminary results show differences in pediatric vaccination rates based on family structure, indicating that children in non-intact families are more likely to receive only mandatory immunisations than those in intact families. To the best of our knowledge, this topic has never been investigated in Italy, a country with a rapidly changing family demography. Exploring determinants of changes in children's vaccination outcomes among non-intact families could be particularly crucial in the realm of public policy, with incomplete vaccination in the pediatric population being identified as a growing public health issue in high-income countries.

Keywords: Non-intact families · pediatric vaccination rates · Italy

1 Background

In the last decades, in many Western countries, the number of separations and divorces has been increasing, resulting in a rise in children and adolescents growing up in non-intact families (Andersson et al., 2017; OECD, 2019). Parental separation affects children's and parents' life domains, changing their lives in many ways. Even if the consequences of divorce vary among the individuals involved (Amato & Anthony, 2014), it is well-documented that union dissolution has several consequences for any children within the couple. Often, children no longer live full-time with both parents, leading to stress and a need for adjustment to the new situation. Moreover, separation brings about a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 617–622, 2025.
https://doi.org/10.1007/978-3-031-64431-3_103

profound transformation in family dynamics, and even in cases of joint residential custody, the non-resident parent (typically the father) often provides less involved parenting, and the parenting style of the resident parent (usually the mother) frequently changes. Moreover, the economic resources of non-intact families tend to worsen (Uunk, 2004; Boertien & Lersch, 2021), affecting (other) children's and parent's life spheres, too. These familial patterns constitute tangible risk factors for children, leading to outcomes that, in alignment with current literature, affect diverse aspects of their lives, ranging from children's educational outcomes (Guetto et al., 2022) to their health and well-being (Kacenelenbogen et al., 2014; Nielsen et al., 2012; Amato, 2006).

Despite the vast amount of literature available on the subject, there is a noticeable lack of comprehensive exploration of separation and divorce patterns and their consequences on parents' attitudes and compliance towards children's vaccination schedules. The existing literature indicated that parental socio-demographic characteristics are crucial in children's vaccination decision-making (Cantuària Tauil et al., 2016). Factors such as parental age (Haynes et al., 2004), parents' education level and income (Anderberg et al., 2011), parents' and siblings' vaccination history and family and household composition (Vandemerulen et al., 2008) have been identified as influential in parents' adherence to vaccination schedules. Moreover, a handful of studies focused on the association between parental separation and children's health, investigating children's vaccination schedules, among other aspects (Vandermeulen et al., 2007; Kacenelenbogen et al., 2014). According to these studies, parental separation is associated with low vaccine coverage and non-compliance with vaccination schedules. As previously discussed, the dynamics of households with two parents may significantly differ from those in separated or single-parent nuclei. Two-parent families might benefit from more shared time, improved recall of vaccination appointments, lower stress levels, and increased involvement in childcare. Separated or divorced couples, and especially non-resident parents, may be less involved in their children's life (Tosi & Guetto, 2024), which could influence their ability to ensure their children receive vaccinations (Kacenelenbogen et al., 2010) and, in general, their engagement in healthcare services. Moreover, union dissolution affects the couple's social resources, degrading the network of individuals (such as grandparents, family friends, common acquaintances, neighbours, etc.) that they can rely on. Therefore, households headed by divorced or separated individuals and single parents could face specific challenges that may influence ordinary children's management of the vaccination calendar.

This work analyses whether family arrangements affect children's pediatric vaccination in Italy, testing if parents heading non-intact families have a lower propensity to vaccinate their children than those with intact families. To the best of our knowledge, this topic has never been investigated in Italy, a country with a rapidly changing family demography.

2 Data and Methods

Data were gathered from an online survey, administered between September 15th and October 10th, 2023, to 12,004 individuals above the age of 40 residing in Italy, and recruited through online panels. The quota sample was stratified by gender (50% men, 50% women), with quotas to match the population's distribution in terms of age class, marital status, education, employment condition, and region of residence. We then restrict our sample to 8,455 respondents who declared to be parents and to have at least one child at the interview date. Our key explanatory variable is a binary variable that identifies those respondents who stopped living together with the partner (because of separation or death) before the youngest child of the couple (or their only child) turned 8 years old and those who never split or who got separated/lived the death of the partner after the youngest child of the couple turned 8. The outcome variable is the youngest child's vaccination outcomes, distinguishing between children who received all or almost all pediatric vaccinations, children who had been administered only compulsory vaccines, and children who did not receive any vaccination.

We estimated a multinomial logistic model for the child's vaccination with the three possible outcomes and the key explanatory variable of whether or not the child has lived in a non-intact family since 8 years old. Moreover, we controlled for some confounders, such as respondents' birth cohort, education level (*Low*: "Up to elementary school", "Middle school", *Medium*: "High school diploma or qualification", *High*: "University degree or bachelor's degree", "Master's degree", "Doctorate or postgraduate specialisation"), and number of children.

3 Preliminary Results

Overall, parents heading a non-intact family before their youngest child turned 8 years old are 724, namely 8.6% of the total sample. Conversely, 7,731 respondents did not separate nor experience the partner's death, at least not before their youngest child turned 8 years old. Among those surveyed, nearly 4% chose not to vaccinate their children, while 45% exclusively adhered to compulsory vaccinations. Approximately 50% of participants ensured their children received all vaccinations outlined in the national plan. Looking at the results of the multinomial regression model presented in Table 1, children living in non-intact nuclei since 8 years old are more likely to receive only compulsory vaccination ($\beta = 0.244$, $P > z = 0.000$) or none vaccination (even if it is only slightly significant at 10%: ($\beta = 0.286$, $P > z = 0.096$) than all pediatric immunisations recommended in the national program, compared to children in two-parent households. As expected, control covariates are in line with the literature, showing, for instance, that low-educated parents as well as parents with three or more children are more likely to intake only compulsory vaccination or none vaccination than all pediatric immunisations.

Table 1. Multinomial regression model estimates, standard errors, and p-values.

Number of children's vaccines (ref = "All vaccines")	Variable	Coeff	Std. Error	P > z
Only compulsory				
	Education (ref = "Medium")			
	Low	0.291	0.055	0.000
	High	− 0.238	0.057	0.000
	Birth cohort (ref = " 1974–1983")			
	1933–1953	0.110	0.068	0.133
	1954–1963	− 0.106	0.065	0.105
	1964–1973	− 0.263	0.063	0.000
	Number of children (ref = "1")			
	2	0.058	0.049	0.234
	3 or more	0.112	0.074	0.074
	Separated since 8 years old (ref = "No")			
	Yes	0.244	0.066	0.000
None				
	Education (ref = "Medium")			
	Low	0.443	0.139	0.001
	High	− 0.080	0.144	0.579
	Birth cohort (ref = "1974–1983")			
	1933–1953	− 1.022	0.170	0.000
	1954–1963	− 1.310	0.175	0.000
	1964–1973	− 0.987	0.147	0.000
	Number of children (ref = "1")			
	2	0.114	0.125	0.364
	3 or more	0.353	0.181	0.051
	Separated since 8 years old (ref = "No")			
	Yes	0.286	0.172	0.096

4 Preliminary Conclusions

In our work, we aim to investigate how children's vaccination decisions are influenced by the family arrangement, in particular, whether the parents live together or whether they stopped living together (because of separation or a parent's death) before the youngest child of the couple (or their only child) turned 8 years old. In this respect, our preliminary analysis showed that children's vaccination outcomes among non-intact families are more likely to be incomplete or even non-existent, thus suggesting that family arrangements play a role in shaping parental behaviour towards children's vaccination.

Exploring the vaccination attitudes of different family structures can provide insights into potential barriers and facilitators to vaccination acceptance. Non-intact families may face unique challenges, such as co-parenting dynamics and conflicting viewpoints between parents, which could influence vaccination decisions differently than two-parent families.

This issue could be particularly crucial in the realm of public policy. In general, incomplete vaccination in the pediatric population has been identified as a growing public health issue in high-income countries, with poorly understood determinants (Jacques et al., 2023). In this context, the substantial increase in divorce rates and the consequential growing share of children living in non-intact families expected in the future would represent a noteworthy aspect in effectively managing vaccination coverage.

References

Amato, P.R.: Marital discord and children's well-being: results from a 20-year longitudinal study of two generations. In: Stewart, C., Dunn, J. (eds.) Families Count: Effects on Child and Adolescent Development, pp. 179–202. Cambridge University Press (2006)

Amato, P.R., Anthony, C.J.: Estimating the effects of parental divorce and death with fixed-effects models. J. Marriage Fam. **76**(2), 370–386 (2014)

Anderberg, D., Chevalier, A., Wadsworth, J.: Anatomy of a health scare: education, income and the MMR controversy in the UK. J. Health Econ. **30**(3), 515–530 (2011)

Andersson, G., Thomson, E., Duntava, A.: Life-table representations of family dynamics in the 21st century. Demogr. Res. **37**(35), 1081–1230 (2017)

Boertien, D., Lersch, P.M.: The long-term consequences of divorce for children's educational attainment in the United Kingdom. Demography **58**(3), 869–895 (2021)

Cantuária Tauil, M., Waldman, E.A., Moura, M.L.D.: Factors associated with incomplete or delayed vaccination across countries: a systematic review. Vaccine **34**(24), 2635–2643 (2016)

Guetto, R., Bernardi, F., Zanasi, F.: Parental education, divorce, and children's educational attainment: evidence from a comparative analysis. Demogr. Res. **46**, 65–96 (2022)

Haynes, K., Stone, C.: Predictors of incomplete immunisation in victorian children. Aust. N. Z. J. Public Health **28**(1), 72–79 (2004)

Jacques, M., et al.: Determinants of incomplete vaccination in children at age two in France: results from the nationwide ELFE birth cohort. Eur. J. Pediatr. **182**(3), 1019–1028 (2023)

Kacenelenbogen, N., Offermans, A.M., Detemmerman, L.: The impact of parental separation on children's health: evidence from Belgium. J. Fam. Stud. **20**(2), 120–134 (2014)

Kacenelenbogen, N., Roland, M., Schetgen, M., Dusart, A.F.: The general practitioner and children of separated parents in Belgium: a qualitative study and its implications. J. Gen. Pract. **1**(131), 2 (2013)

Nielsen, N.M., Hansen, A.V., Simonsen, J., Hviid, A.: Stressful life events in childhood and risk of infectious disease hospitalization. Eur. J. Pediatr. **171**, 173–179 (2012)
OECD: Society at a Glance 2019: OECD Social Indicators. OECD Publishing, Paris (2019)
Tosi, M., Guetto, R.: The social stratification in parent-child relationships after separation: evidence from Italy. J. Fam. Res. **36**, 25–42 (2024)
Uunk, W.: The economic consequences of divorce for women in the European Union: the impact of welfare state arrangements. Eur. J. Popul. **20**(3), 251–285 (2004)
Vandermeulen, C., Roelants, M., Theeten, H., Van Damme, P., Hoppenbrouwers, K.: Vaccination coverage and sociodemographic determinants of measles–mumps–rubella vaccination in three different age groups. Eur. J. Pediatr. **167**, 1161–1168 (2008)

Quantile Treatment Effect and GBM: An Approach Based on Propensity Score

Francesco Dotto[1(✉)], Francesco Giuli[1], Margherita Scarlato[1], and Francesco Bloise[2]

[1] Roma Tre University, Via Silvio D'Amico, 77., 00145 Rome, Italy
francesco.dotto@uniroma3.it
[2] Sapienza University, via del Castro Laurenziano, 9., 00161 Rome, Italy

Abstract. We explore how the use of the Gradient Boosting Machine (GBM) method to compute propensity scores may improve the estimation of the Quantile Treatment Effect (QTE) in the case of a binary treatment and a high dimensional dataset. To validate the procedure we provide a simulation study on several scenarios and apply the method to estimate the wage gap between workers in the informal and formal sectors in South Africa at different quantiles of the wage distribution.

Keywords: QTE · GBM · Causal Inference

1 Introduction and Methodology

Propensity score analysis is widely used in quasi-experimental studies to address sample selection bias issues, compare [1], and many examples exist in several applied contexts such as the field of labour economics, where the dimension of confounders is usually very high. Recently, the usage of machine learning algorithms have been proposed as a strategy to deal with this challenge. An example has been implemented by [2], who show that machine learning algorithms for estimating the propensity scores provide more credible results of the Average Treatment Effect on the Treated (ATET) when estimating the effect of a training program in Germany using a high-dimensional dataset. In this contribution, we contribute to this emerging literature by exploring how machine learning methods to compute the propensity scores can improve the estimation of the Quantile Treatment Effect (QTE) in the case of a binary treatment. We then provide a real data application of the proposed method to estimate the informal-formal wage gaps across the wage distribution in the South Africa' s labour market.

Methodologically speaking, we follow the potential outcome approach as in [3] and in [4] and denote by $i = 1, 2, \ldots, n$ the i-th statistical unit and by T_i the treatment assignment taking value 1 if the participant i is treated and 0 if the i-th participant is not treated. Furthermore let $Y_i(1)$ denote the potential outcome when the i-th participant is treated and $Y_i(0)$ the potential outcome

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 623–628, 2025.
https://doi.org/10.1007/978-3-031-64431-3_104

when the i-th participant is not treated. Such outcomes cannot be observed simultaneously and we can only observe the i-th outcome given by $Y_i = T_i Y_i(1) - (1 - T_i)Y_i(0)$. The goal of our analysis is to assess the wage gap between workers employed in the informal and formal sector at each selected quantile of the wage distribution. Methodologically speaking, to assess the issue outlined above, one needs to evaluate the QTE at each quantile of interest $q \in (0, 1)$

$$QTE(q) = F^{-1}_{Y(1)}(q) - F^{-1}_{Y(0)}(q) \tag{1}$$

where, in Eq. (1), $F^{-1}_{Y(1)}$ and $F^{-1}_{Y(0)}$ identify the quantile function of the treated and untreated units which, in our case, are given by workers who are employed in the informal sector and workers who are employed in the formal sector, respectively, while the outcome variable Y is given by the hourly net wage. In order to balance treated and untreated sample units, as already done in [1] and in [5] for the specific case of the QTE, we use a propensity score weighted quantile regression where the propensity scores are obtained by modelling the interaction of the treatment variable T and set of available covariates $\boldsymbol{X}$, where T, in our case, is the variable denoting the status of being employed in the informal sector and Y indicates the wage. As workers are sampled with unequal probabilities, the effect of informality can not be estimated directly and each subject must be weighted according to the joint distribution of the treatment variable and the covariates.

2 Data and Simulation Study

For the empirical analysis we use the NIDS survey and focus on the last wave which has been carried out in 2017. NIDS is a nationally representative survey conducted by the South African Labour and Development Research Unit (SALDRU) of the University of Cape Town. The aim of the survey is to follow a sample of household members and register changes in household compositions and migrations and several dimensions of well-being (such as incomes, expenditures, assets, access to social services, education, health, and employment).[1]. After properly manipulating and subsetting the data[2], our dataset is composed of $n = 6,474$ observations and $p = 43$ variables.

To validate the proposed approach, we run a simulation study. The aim of this section is to compare the efficiency in estimating the QTE at five quantile levels 0.10, 0.25, 0.50, 0.75 and 0.90. As in [5], these estimates are obtained by means of a weighted quantile regression model. More in details, we use several approaches for estimating the probability of treatment, $\mathbb{P}[T = 1|\boldsymbol{X}]$, which are fully described below. Then, in order to compute the QTE, we run, for each

[1] For a detailed description of the data compare, among the others, [6].

[2] The data can be easily obtained by downloading the whole dataset and by manipulating it by using the STATA macro available on the data repository web site: http://www.nids.uct.ac.za/nids-data/program-library/derived-files.

generated dataset, a quantile regression model that includes T as a covariate. Within those models, each observation is weighted with weights computed as

$$\frac{T_i}{\mathbb{P}[T_i = 1|\boldsymbol{X}_i]} + \frac{(1 - T_i)}{(1 - \mathbb{P}[T_i = 1|\boldsymbol{X}_i])} \tag{2}$$

The simulated data are composed by $n = 5,000$ data points, $p = 12$ covariates, both continuous and categorical, a dichotomous treatment variable T and a continuous outcome variable $Y = T_i Y_i(1)_i + (1 - T_i)Y_i(0)$ with $Y(0) \sim N(7, 5)$ and $Y(1) \sim N(13, 26)$. The following procedures are compared within the simulation study:

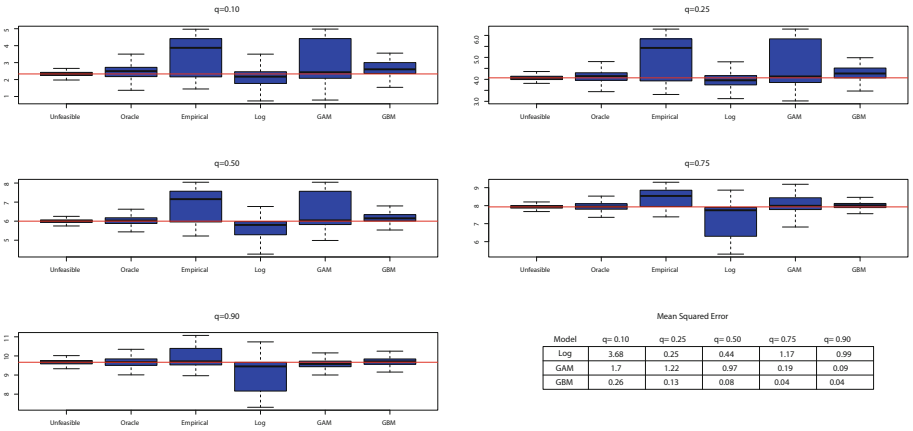
- "Unfeasible estimator" (labelled as "Unfeasible"). Such estimator, which is obviously unfeasible in reality, is based on computing the QTE by using the quantity $QTE(q) = F^{-1}_{Y_1}(q) - F^{-1}_{Y_0}(q)$ which is observable only on simulated data as, for each observation i we either observe $Y(1)_i$ or $Y(0)_i$. Thus, such estimator has to be considered as a kind of "golden standard" whose performances cannot be reached as in the real data we can only observe $Y_i(1)$ or $Y_i(0)$;
- "Oracle estimator" (labelled as "Oracle"). This estimator computes the QTE by using a weighted quantile regression model. The involved weights are obtained involving the true probabilities $\mathbb{P}[T = 1|\boldsymbol{X}]$;
- "Empirical estimator". This estimator uses only the observed outcome to compute the QTE;
- "Logistic model estimator" (labelled as "Log"). This estimator is similar to the previous one except for the propensity scores which are computed from probabilities estimated after implementing a logistic model;
- "Generalized additive logistic model estimator" (labelled as "GAM"). Such estimator is the one adopted in [5] and is based on the methodology presented in [7];
- "Gradient boosting estimator" (labelled as "GBM"). In this case the treatment probabilities are computed by using a GBM model.

In the three statistical models (Log, GAM and GBM), the 12 covariates available are included. This is done in order to assess how the different approaches deal with potentially irrelevant variables.

For the sake of brevity, we report in Fig. 1, the aggregated results of the simulation study

3 Empirical Analysis: Modelling the Effect of Informality

Within this section we report the results of the quantile regression model in which the outcome variable is the net hourly wage. As we focus on the percentage of the reduction of the salary due to the status of informal worker, we divide the intercept of the model at each quantile by the corresponding slope.



Model	q= 0.10	q= 0.25	q= 0.50	q= 0.75	q= 0.90
Log	3.68	0.25	0.44	1.17	0.99
GAM	1.7	1.22	0.97	0.19	0.09
GBM	0.26	0.13	0.08	0.04	0.04

Fig. 1. Simulation results (aggregation of all scenarios)

Table 1 reports the estimated QTE adopting the three different classification models and the empirical QTE and the associated standard errors. It shall be remarked that, as we consider the ratio among two parameters, there is no closed form for the standard errors which are computed in two steps. In the first one, we apply the bootstrap to obtain an estimate of the standard error ([8]) to each parameter and then we approximate the value of the standard error of the ratio via the delta method. The regression-based models and the empirical approach yield similar results. According to such estimates, the effect of informality causes a wage penalty that ranges from to 40% in the left part of the earnings distribution to 60% in the right part of the distribution. The GBM-IPW estimator provides different results. Indeed, the effect of informality on wages along the distribution is still negative, confirming the hypothesis of segmentation and rigidities in the labor market. In addition the wage penalty varies along the wage distribution and workers attributes, but in this case the lower-tier informal jobs are the most penalized. Interestingly, the penalty at each quantile of interest is significantly reduced with respect to the other estimators, suggesting that this method is able to account for the individual characteristics of workers more precisely. In facts, the prediction accuracy provided by the GBM is significantly higher and furthermore, differently from regression-based methods that need to work on the complete cases (observations for which values are available on all the covariates), GBM treats the missing values as a separate category allowing the insertion of such observations in the model.

Table 1. Wage penalty of informal workers estimated via quantile regression

Procedures	q_{10}	q_{25}	q_{50}	q_{75}	q_{90}
Empirical	-0.444	-0.389	-0.444	-0.62	-0.619
	(0.022)	(0.019)	(0.012)	(0.015)	(0.025)
Log	-0.527	-0.429	-0.444	-0.615	-0.612
	(0.042)	(0.038)	(0.033)	(0.034)	(0.046)
GAM	-0.467	-0.435	-0.475	-0.653	-0.663
	(0.033)	(0.032)	(0.033)	(0.033)	(0.057)
GBM	-0.289	-0.216	-0.206	-0.266	-0.259
	(0.044)	(0.024)	(0.031)	(0.053)	(0.071)

Notes: The quantile regression models are weighted using the propensity scores obtained by means of Log, GAM and GBM models. The empirical procedure consists in comparing the wages of formal and informal workers at each quantile without weighting the observations. In brackets we report the standard errors of the parameter estimates.

4 Conclusions

In this contribution we outlined a novel approach for computing the weighted QTE using the propensity scores estimated by the GBM method. Even if the methodology is not new in the statistical and econometrics literature, it has never been applied in such context at our best knowledge. We carry out a simulation study to compare this method with regression-based methods used in the literature and then provide a real data application to the estimation of the informal sector wage gap across the wage distribution in the case of the South Africa's labor market. Also in this case, our paper provide a novel contribution addressing the selection on observable issue when estimating the informal wage penalties across the wage distribution. A limit of the proposed method is the one that time dependent data cannot be analyzed as machine learning method only work with independent observations.

References

1. 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)
2. Goller, D., Lechner, M., Moczall, A., Wolff, J.: Does the estimation of the propensity score by machine learning improve matching estimation? the case of Germany's programmes for long term unemployed. Labour Econ. **65**, 101855 (2020)
3. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688–701 (1974)

4. Rubin, D.B.: Causal inference using potential outcomes: design, modeling, decisions. J. Am. Stat. Assoc. **100**(469), 322–331 (2005)
5. Firpo, S.: Efficient semiparametric estimation of quantile treatment effects. Econometrica **75**, 259–276 (2007)
6. Leibbrandt, C.M., Woolard, I.: National income dynamics study panel user manual. Cape Town South. Africa Labour Dev. Res. Unit (2016)
7. Hastie, T., Tibshirani, R.: Generalized additive models: some applications. J. Am. Stat. Assoc. **82**(398), 371–386 (1987)
8. DiCiccio, T.J., Efron, B.: Bootstrap confidence intervals. Stat. Sci. **11**(3), 189–228 (1996)

High Volatility, High Emissions? A Hidden-Markov Model Approach

Pierdomenico Duttilo[1(✉)], Marina Bertolini[1], Alfred Kume[2], and Stefano Antonio Gattone[3]

[1] University of Padova, DSS, Via C. Battisti 241, 35121 Padova, Italy
pierdomenico.duttilo@unipd.it
[2] University of Kent, SMSAS, Park Wood Road, CT2 7FS,Canterbury, UK
[3] University "G. d'Annunzio", DISFIPEQ, Viale Pindaro 42, 65127 Pescara, Italy

Abstract. This study introduces the generalised normal distribution hidden Markov model (GND-HMM) with constrained and unconstrained parameters, where a first-order Markov chain governs the draws of the states from the mixture components. The proposed model is applied to the daily electricity price returns of the electricity market in northern Italy to detect the turmoil periods that occurred during the years 2020–2023. The turmoil periods detected by the GND-HMM model highlight important aggregate events such as the Covid-19 pandemic and the Russia-Ukranian conflict. Furthermore, our study aims to examine the relationship between the identified turmoil periods and the time series of CO_2 emissions in the northern Italian electricity market.

Keywords: hidden Markov models · electricity prices · CO_2 emissions

1 Introduction

Due to the rise in geopolitical uncertainties around the world, there have been significant fluctuations in electricity prices throughout Europe [1]. In electricity markets, price volatility, especially when detected in specific market zones, might be heavily influenced by the infrastructures and the energy mix of the area. Therefore, periods of high volatility in the electricity market can result in changes in the generation of electricity at the margin. For example, during periods of high demand or supply disruptions, there may be a greater dependence on certain types of power plants (such as coal or oil plants) that emit higher levels of CO_2 compared to cleaner sources such as natural gas or renewables [2]. Analysing turmoil periods in the energy market is crucial to predict its future changes and might be relevant to understand its implications in terms of CO_2 emissions.

In this paper, we introduce the generalised normal distribution hidden Markov model (GND-HMM) with constrained and unconstrained parameters, applying it to the Italian electricity market and, in particular, to the day ahead market for the northern part of Italy. We want to start analysing the day ahead

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 629–634, 2025.
https://doi.org/10.1007/978-3-031-64431-3_105

market even if it is less exposed to sudden changes in production, but further analysis on balancing markets will be carried out in future research.

After this brief introductory section, we turn to Sect. 2 where the GND-HMM is introduced. The results are presented in Sect. 3. Finally, some conclusions are provided in Sect. 4.

2 Data and Methodology

The study is carried out considering the daily electricity prices of the day-ahead market of northern Italy ($\text{MGP}_{\text{North}}$). Most electricity transactions occur within the MGP, where prices and demand for the following day are established through simultaneous auctions every hour. Data on daily electricity prices are collected from Gestore Mercati Energetici (GME) for the period from January 1, 2020 to January 31, 2024 [3]. The daily returns are computed as follows

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right) = \log(P_t) - \log(P_{t-1}), \tag{1}$$

where r_t denotes the log return on electricity price a time t, P_t is the electricity price at time t, while P_{t-1} is the electricity price at time $t-1$.

A random variable X_t is said to have a K-component finite mixture of generalized normal distributions (MGND) if its density is given by

$$f(x_t|\pi_1, ...\pi_K, \theta_1, ..., \theta_K) = \sum_{k=1}^{K} \pi_k f_k(x_t|\theta_k), \tag{2}$$

where $0 < \pi_k < 1$, $k = 1, ..., K$, $\sum_{k=1}^{K} \pi_k = 1$ are the mixture weights, and $f_k(x_t|\theta_k)$, $k = 1, ..., K$, are the component densities taken to be generalized normal distributions (GND),

$$f_k(x_t|\theta_k) = f_k(x_t|\mu_k, \sigma_k, \nu_k) = \frac{\nu_k}{2\sigma_k \Gamma(1 \,/\, \nu_k)} \exp\left\{-\left|\frac{x_t - \mu_k}{\sigma_k}\right|^{\nu_k}\right\} \tag{3}$$

where $\Gamma(1 \,/\, \nu_k) = \int_0^\infty t^{1/\nu_k - 1} \exp^{-t} dt$, μ_k is the k-th location parameter ($\mu_k \in$ R), σ_k is the k-th scale parameter ($\sigma_k > 0$), and ν_k is the k-th shape parameter ($\nu_k > 0$). Figure 1 shows the k-th probability density function (Eq. 3) for $\mu_k = 1$, $\sigma_k = 1$. The shape parameter ν_k controls both the peakedness and the tail weights. If $\nu_k = 1$ the GND reduces to the Laplace distribution and if $\nu_k = 2$ it coincides with the normal distribution. It is observed that $1 < \nu_k < 2$ produces an "intermediate distribution" between the normal distribution and the Laplace distribution. As limit cases, for $\nu \to \infty$ the distribution tends to a uniform distribution, while for $\nu \to 0$ it will be impulsive [4].

Wen et al. (2022) studied a two-component mixture of univariate GND and proposed an expectation conditional maximisation (ECM) algorithm for parameter estimation [4]. Modelling the return density using the MGND model (Eq. 2) is

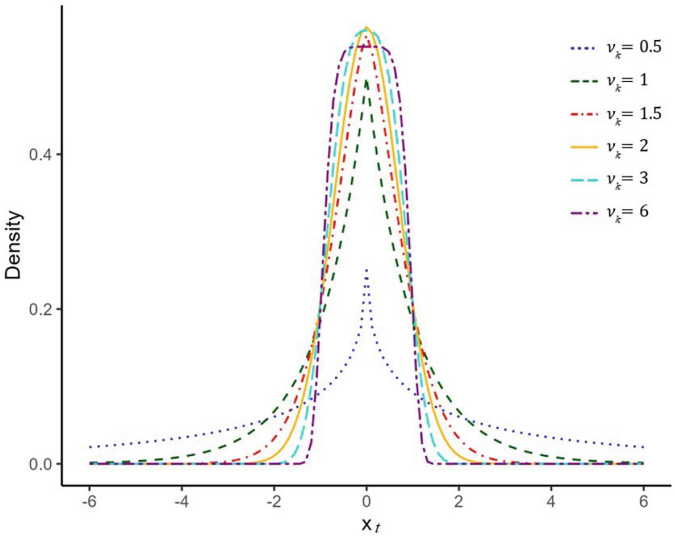


Fig. 1. The k-th probability density function for $\mu_k = 0$ and $\sigma_k = 1$.

advantageous for two main reasons. First, it is a flexible tool capable of capturing the skewness and excess kurtosis of daily returns [4]. Secondly, from a theoretical perspective, the existence of various mixture components can be attributed to several market states (or regimes). Usually, financial markets are characterised by two market periods: stable and tumultuous. The latter are caused by aggregate events like economic and geopolitical shifts. Since the mixture components could be different with respect to their variances, a daily return belongs to the stability period if it belongs to the mixture component with the lowest variance. Similarly, a daily return belongs to the turmoil period if it belongs to the mixture component with the highest variance.

However, as already identified by Mandelbrot 1963, daily returns exhibit volatility clustering, showing small fluctuations during stable periods and large fluctuations during turmoil periods [5,6]. In statistical terms, this means heteroskedasticity, i.e., non-constant variance over time. Independent mixture models, such as the MGND model, cannot fully capture heteroskedasticity, since they do not account for temporal dependence [7]. Therefore, it is necessary to define the stochastic process that determines the market states (or regimes), which involves changes in the mixture weights π_k for $k = 1, ..., K$. The mixture mechanism should be set in a way that emphasises the probability of choosing the same mixture component at time $t + 1$ as the one selected at time t.

In such situations, Markov chains become essential. Originally, hidden Markov models (HMM) were introduced by Baum and Petrie (1966) with their landmark paper titled "Statistical Inference for Probabilistic Functions of Finite State Markov Chains" [8]. A sequence of discrete random variables $\{S_t : t \in \mathbb{N}\}$ is a Markov chain if for all $t \in \mathbb{N}$ the Markov property is sat-

isfied $\Pr(S_{t+1}|S_t, ..., S_1) = \Pr(S_{t+1}|S_t)$. The future state S_{t+1} depends only on its current state S_t and not on the state sequence that preceded it [9]. A Markov chain is represented by important quantities called transition probabilities $\gamma_{ij}(t) = \Pr(S_{c+t} = j|S_c = i)$. An homogeneous ($\gamma_{ij}(t) = \gamma_{ij}$ for all t) Markov chain with K states can be defined by a $K \times K$ transition probability matrix (t.p.m.)

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \ddots & \vdots \\ \gamma_{K1} & \cdots & \gamma_{KK} \end{pmatrix}, \text{ with } \gamma_{ij} \in [0,1] \text{ for all } ij \text{ and } \sum_{k=1}^{K} \gamma_{ij} = 1. \tag{4}$$

A Markov chain with t.p.m. $\mathbf{\Gamma}$ have stationary distribution $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)'$ if $\boldsymbol{\pi}\mathbf{\Gamma} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{1}' = 1$. The vector $\boldsymbol{\pi}$ is subject to $\sum_{k=1}^{K} \pi_k = 1$ with $\pi_k > 0$. Each element π_k represents the probability that the chain is in state k when the chain is in a state described by $\boldsymbol{\pi}$. The MGND model (Eq. 2) together with the Markov chain (Eq. 4) gives the GND-HMM.

In addition, to obtain more parsimonious models, constraints are imposed on the location, scale, and shape parameters: $\mu_k = \mu$, $\sigma_k = \sigma$, $\nu_k = \nu$, for $k = 1, ..., K$. With $K = 2$, taking into account all possible combinations of these constraints would result in a family of eight models. Parameter estimation is performed with the direct optimisation method proposed by Zucchini et al. [9]. The likelihood function of the GND-HMM is given by

$$L_T = \Pr\left(\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}\right) = \boldsymbol{\pi}\boldsymbol{P}(x_1)\mathbf{\Gamma}\boldsymbol{P}(x_2)\ldots\mathbf{\Gamma}\boldsymbol{P}(x_T)\mathbf{1}', \tag{5}$$

where $\boldsymbol{X}^{(T)} \equiv (X_1, X_2, ..., X_T)$, $\boldsymbol{\pi}$ is the initial distribution of S_1 and $\boldsymbol{P}(x)$ is the $K \times K$ diagonal matrix with diagonal element equal to $f_k(x_t|\mu_k, \sigma_k, \nu_k)$, the k-th state-dependent distribution.

3 Results

Table 1 shows the two-state GND-HMM estimates with and without constraints for the daily returns on $\text{MGP}_{\text{North}}$. Models belonging to the 8-model family are compared with the Bayesian Information Criterion (BIC). The number of estimated parameters, $\dim(\hat{\theta})$, varies by moving from the UUU model to the CCC model. According to the BIC, the CCU is the best model, as it reports the lowest BIC (-1627.51). Figure 2 panel (a) shows the estimated marginal distribution of the selected model. It can be seen that the predominant stable component (dotted line) describes the central and intermediate values of the data, while the turmoil component (dashed line) describes the tail behaviours. The shape parameter of the turmoil component ($\hat{\nu}_2 = 0.9637$) is lower than the shape parameter of the stable component ($\hat{\nu}_1 = 1.5966$). Specifically, the turmoil component has heavier tails than a Laplace distribution. Figure 2 panel (b) shows the decoded turmoil periods (coloured yellow) using the Viterbi algorithm [10].

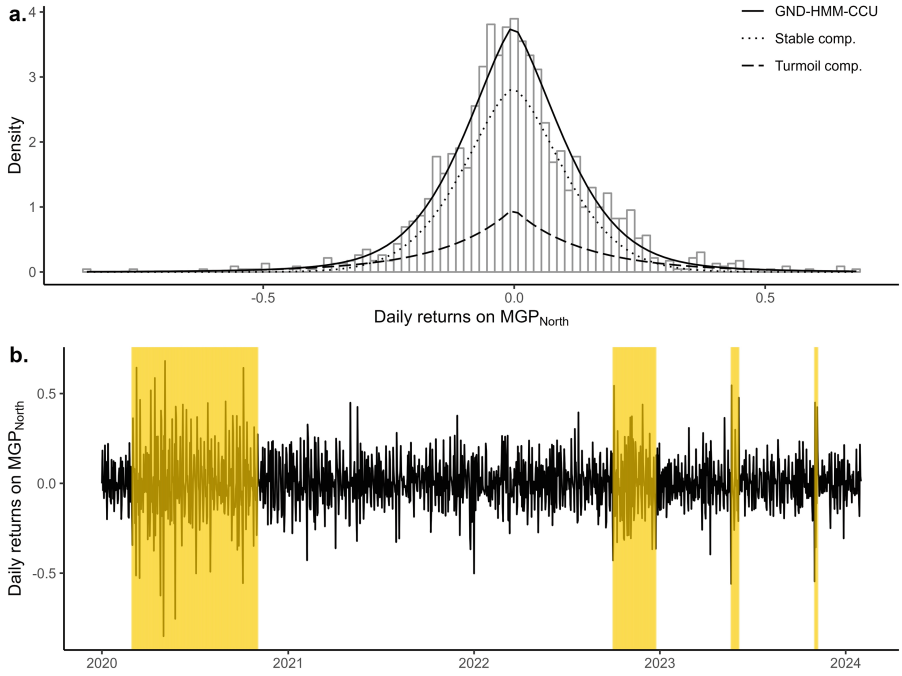


Fig. 2. Estimated marginal distribution in panel (a) and turmoil periods (coloured yellow) decoded through the Viterbi algorithm in panel (b).

Table 1. Estimated two-state GND-HMM models for daily returns on $\text{MGP}_{\text{North}}$.

	GND-HMM							
θ	UUU	CUU	UCU	UUC	CCU	CUC	UCC	CCC
π_1	0.7190	0.7217	0.7099	0.7382	0.7172	0.7375	0.6880	0.5000
μ_1	-0.0003	-0.002	0.0002	-0.0017	-0.0022	-0.0024	-0.0070	-0.0050
σ_1	0.1378	0.1366	0.1428	0.1256	0.1420	0.1256	0.1209	0.1211
ν_1	1.5549	1.5360	1.6215	1.3877	1.5966	1.3890	1.0927	1.0943
π_2	0.2810	0.2783	0.2901	0.2618	0.2828	0.2625	0.3120	0.5000
μ_2	-0.0108	-0.002	-0.0113	-0.0078	-0.0022	-0.0024	-0.0004	-0.0050
σ_2	0.1734	0.1787	0.1428	0.2324	0.1420	0.2327	0.1209	0.1211
ν_2	1.0929	1.1147	0.9700	1.3877	0.9637	1.3890	1.0927	1.0943
γ_{11}	0.9907	0.9905	0.9911	0.9886	0.9915	0.9883	0.6166	0.8000
γ_{22}	0.9762	0.9753	0.9783	0.9678	0.9784	0.9672	0.1546	0.8000
$\dim(\hat{\theta})$	9	8	8	8	7	7	7	6
$\log L(\hat{\theta})$	840.87	840.31	840.16	838.33	839.33	838.15	802.64	802.59
BIC	-1615.97	-1622.15	-1621.86	-1618.19	**-1627.51**	-1625.15	-1554.13	-1561.33

Notes. In bold the best model according to the BIC.

4 Final Remarks

This study introduces GND-hidden Markov models with unconstrained and constrained parameters, where a first-order Markov chain governs the draws of the states of the mixture components. The proposed model is applied to the daily electricity price returns of the electricity market in northern Italy ($\text{MGP}_{\text{North}}$) to detect recent turmoil periods. The detected turmoil periods in Fig. 2 panel (b) highlight important aggregate events such as the outbreak of the Covid-19 pandemic and the Russia-Ukranian conflict.

Future research will focus on finding out whether the presence of turbolences in the zonal market corresponds to a higher use of pollutant energy sources or vice versa if they represent higher penetration of renewables. In this sense, we want to analyse the trade-off between market stability and the level of emissions. Future extensions will take into account data from all the Italian geographical zones for the MGP market and, later, on the national balancing market.

Acknowledgments. This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C93C22005270001). 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. Liu, D., Liu, X., Guo, K., Ji, Q., Chang, Y.: Spillover effects among electricity prices, traditional energy prices and carbon market under climate risk. Int. J. Environ. Res. Public Health **20**, 1–18 (2023)
2. Sovacool, B.K.: The intermittency of wind, solar, and renewable electricity generators: technical barrier or rhetorical excuse? Util. Policy **17**, 288–296 (2009)
3. Gestore dei Mercati Energetici (GME). https://www.mercatoelettrico.org/it/
4. Wen, L., Qiu, Y., Wang, M., Yin, J., Chen, P.: Numerical characteristics and parameter estimation of finite mixed generalized normal distribution. Commun. Stat. Simul. Comput. **51**, 3596–3620 (2022)
5. Mandelbrot, B.: A subordinated stochastic process model with finite variance for speculative prices: comment. Econometrica **41**, 157–59 (1973)
6. Haas, M. Pigorsch C.: Financial economics, fat-tailed distributions. Encycl. Complex. Syst. Sci. 3404–3435 (2009). Springer New York
7. Kaehler, J., Marnet, V.: Markov-switching models for exchange-rate dynamics and the pricing of foreign-currency options. In: Econometric Analysis of Financial Markets, pp. 203–230 (1994). Heidelberg: Physica-Verlag HD.
8. Baum, L. E. and T. Petrie : Statistical inference for probabilistic functions of finite state markov chains. Ann. Math. Stat. **37**, 1554–1563 (1966)
9. Zucchini, W., MacDonald, I., Langrock. R.: Hidden Markov Models for Time Series: An Introduction Using R, Second Edition. Chapman & Hall. CRC Press (2017)
10. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inform. Theory **13**, 260–269 (1967)

Mixtures of Generalized Latent Trait Analyzers for Biclustering Multivariate Data

Dalila Failli[1(✉)], Maria Francesca Marino[1], and Francesca Martella[2]

[1] Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni, Viale Morgagni 59, 50134 Firenze, Italy
dalila.failli@unifi.it

[2] Sapienza Università di Roma, Dipartimento di Scienze Statistiche, Piazzale Aldo Moro 5, 00185 Roma, Italy

Abstract. A novel model-based biclustering approach for multivariate data is introduced exploiting a finite mixture of generalized latent trait models. The proposed model clusters units into distinct subsets, called components. Within each component, subsets of variables, called segments, are identified by specifying the linear predictor in terms of a row-stochastic vector. The continuous latent trait integrated into the model allows us to account for the residual dependence between multivariate outcomes from the same unit. We employ an EM algorithm for maximum likelihood estimation of model parameters, with Gauss-Hermite quadrature utilized to approximate multidimensional integrals where closed-form solutions are not available.

Keywords: Model-based clustering · GLMs · Dependent observations · EM algorithm · Gaussian quadrature

1 Introduction

Biclustering [2,8–10] aims at concurrently partitioning units and variables into homogeneous blocks of rows and columns within a data matrix. This approach is often used to analyze large data matrices where the relationships between rows and columns can be considered as symmetrical. Literature on biclustering can be broadly divided into exploratory and model-based approaches. Here, we focus on the latter and extend the Mixture of Latent Trait Analyzers (MLTA) originally proposed by [7] to general multivariate responses with respect to distribution in the Exponential Family (EF; [3,13]). The original MLTA model is specifically designed for categorical (binary) data and enables clustering of units via a finite mixture specification. In this paper, we extend the proposal in two directions. Firstly, we enable the model to handle both qualitative and quantitative variables. Secondly, following the approach outlined in [5], we allow for a joint clustering of units and variables, where units are partitioned into subsets called

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 635–639, 2025.
https://doi.org/10.1007/978-3-031-64431-3_106

components and, within each component, variables are partitioned into subsets called segments. This goal is achieved by considering a finite Mixture of Generalized Latent Trait Analyzers (MGLTA) providing a partitioning of units, and a flexible specification of the linear predictor for partitioning variables, drawing upon an approach akin to the proposal outlined in [1,12]. The relationship between variables is modeled via a continuous, latent trait, similar to the original MLTA. Moreover, following the suggestion in [6], we introduce the possibility of incorporating covariates into the latent layer of the model to ascertain their impact on component formation.

2 The MLTA Model

Let Y_{ik} denote the k-th binary random variable and y_{ik} the corresponding observed value for the i-th unit, with $i = 1, \ldots, N$, and $k = 1, \ldots, R$. In the context of categorical (binary) outcomes, the MLTA model assumes that each unit is assigned to one out of G unobserved components, identified by a discrete latent variable $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})' \overset{iid}{\sim} \text{Multinomial}(1, (\eta_1, \ldots, \eta_G)')$. Each element of this vector, z_{ig}, equals 1 if unit i belongs to component g, and 0 otherwise. The parameter η_g represents the probability that a randomly selected unit belongs to component g, subject to the constraints $\sum_{g=1}^{G} \eta_g = 1$ and $\eta_g \geq 0$, for $g = 1, \ldots, G$. Furthermore, the MLTA assumes the existence of a D-dimensional, unit-specific, continuous latent trait $\boldsymbol{u}_i$, distributed according to a standard D-variate Gaussian density, i.e., $\boldsymbol{u}_i \sim \mathcal{N}_D(\boldsymbol{0}, \boldsymbol{I})$. This assumption aims to capture the heterogeneity between units concerning the observed outcomes. Given $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$, the responses in the vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iR})'$ are assumed to be independent Bernoulli random variables with probabilities $\pi_{gk}(\boldsymbol{u}_i)$, $k = 1, \ldots, R$, and these probabilities are modeled through the following logistic function:

$$\text{logit}(\pi_{gk}(\boldsymbol{u}_i)) = b_{gk} + \boldsymbol{w}'_{gk}\boldsymbol{u}_i,$$

where b_{gk} represents the baseline tendency of units in the g-th component to respond to the k-th item, and $\boldsymbol{w}_{gk}$ measures the heterogeneity between units in the g-th component in the way they respond to the observed items.

3 The MGLTA Model

Let us suppose that Y_{ik} represents the k-th response associated to unit i, and y_{ik} the corresponding realization. Both qualitative and quantitative variables are allowed. As before, let $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})'$ be a Multinomial random variable, allowing the partitioning of units into one of G different components. Differently from the original MLTA model, we allow for unequal component probabilities across units, as suggested by [6]. In detail, we consider a multinomial logit specification of the type:

$$\eta(\boldsymbol{x}_i; \boldsymbol{\beta}_g) = \Pr(z_{ig} = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}_g) = \frac{\exp\{\boldsymbol{x}'_i\boldsymbol{\beta}_g\}}{1 + \sum_{g'=2}^{G} \exp\{\boldsymbol{x}'_i\boldsymbol{\beta}_{g'}\}}, \quad g = 2, \ldots, G,$$

where $\boldsymbol{\beta}_g$ denotes a J-dimensional vector of coefficients measuring the impact of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})'$ on $\eta(\boldsymbol{x}_i; \boldsymbol{\beta}_g)$. In addition, let $\boldsymbol{u}_i \sim \mathcal{N}_D(\mathbf{0}, \boldsymbol{\Sigma})$ be a D-dimensional, continuous, latent trait, following a multivariate Gaussian density with zero mean vector and covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$. We assume that, conditional on $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$, the random responses Y_{ik} are independent each other, with density in the Exponential Family (EF). Consequently, the joint (conditional) density of the vector $\boldsymbol{y}_i$ may be expressed as:

$$\begin{aligned} f(\boldsymbol{y}_i \mid \boldsymbol{u}_i, z_{ig} = 1) &= \prod_{k=1}^{R} f(y_{ik} \mid \gamma_{gk}(\boldsymbol{u}_i), \alpha_{gk}) \\ &= \prod_{k=1}^{R} \exp\left\{ \frac{y_{ik}\gamma_{gk}(\boldsymbol{u}_i) - c(\gamma_{gk}(\boldsymbol{u}_i))}{b(\alpha_{gk})} + d(y_{ik}; \alpha_{gk}) \right\}, \end{aligned}$$

where $\gamma_{gk}(\boldsymbol{u}_i)$ and α_{gk} represent the canonical parameter and the dispersion parameter of the EF, respectively. Further, $b(.)$, $c(.)$, and $d(.)$ denote known functions. Within each component, to identify variable segments (i.e., partitions of variables), we adopt a similar approach to that proposed by [5]. We assume that the canonical parameter, $\gamma_{gk}(\boldsymbol{u}_i)$, is described by the following equation:

$$\gamma_{gk}(\boldsymbol{u}_i) = b_g + \boldsymbol{a}'_{gk}(\boldsymbol{\mu} + \boldsymbol{u}_i), \tag{1}$$

where b_g is a component-specific effect providing a baseline (transformed) mean value for units in the g-th component, while $\boldsymbol{\mu}$ is a D-dimensional vector of fixed parameters assumed to be constant across components. Lastly, $\boldsymbol{a}_{gk}$ is a D-dimensional row stochastic vector ($D \leq R$) with elements:

$$a_{gkd} = \begin{cases} 1 & \text{if, within the } g\text{-th component, the } k\text{-th variable belongs to segment } d, \\ 0 & \text{otherwise.} \end{cases}$$

This latter allows for the selection of one out of the D elements in the vectors $\boldsymbol{\mu}$ and $\boldsymbol{u}_i$. Consequently, for a unit in the g-th component and a variable in the d-th segment, the canonical parameter is modeled as:

$$\gamma_{gk}(u_{id}) = b_g + a_{gkd}(\mu_d + u_{id}).$$

We may note that this specification allows for a different number of variable-specific segments in each mixture component, allowing the model to better capture the underlying patterns in the data.

3.1 Parameter Estimation and Inference

Let $\boldsymbol{\theta}$ represent the vector of all free model parameters. The likelihood function of the model can be written as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left(\sum_{g=1}^{G} \eta(\boldsymbol{x}_i; \boldsymbol{\beta}_g) \int_{R^D} f(\boldsymbol{y}_i \mid \boldsymbol{u}_i, z_{ig} = 1) f(\boldsymbol{u}_i) d\boldsymbol{u}_i \right).$$

Maximum likelihood estimates of model parameters can be obtained by maximizing the log-likelihood function via an EM algorithm [4]. The E-step of this algorithm involves solving multiple integrals that do not have a closed-form solution. We propose to approximate these integrals using Gauss-Hermite quadrature [15,18]. An initialization strategy based on multiple starting points can help the estimation algorithm in avoiding local maxima of the likelihood function. At convergence of the algorithm, each unit can be assigned to the g-th component using a Maximum a Posteriori (MAP) rule based on the estimated posterior probabilities $\hat{z}_{ig}$, while each variable is automatically assigned to one out of the D segments according to the vector $\hat{\boldsymbol{a}}_{gk}$.

3.2 Standard Errors and Model Selection

Standard errors for the estimates derived from the EM algorithm can be obtained by following the method suggested by [11,14], and by considering a sandwich formula [16,19]. In detail, they correspond to the square root of the diagonal elements of the following covariance matrix:

$$\widehat{Cov}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1},$$

where $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})$ is the observed information matrix, obtained by computing the first numerical derivative of the score vector $S(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, evaluated at $\hat{\boldsymbol{\theta}}$. In addition, $\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})$ is an estimate for the covariance matrix of the score vector and is computed as $\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} S_i(\hat{\boldsymbol{\theta}})S(\hat{\boldsymbol{\theta}})'$, where $S_i(\hat{\boldsymbol{\theta}})$ denotes the individual contribution to the score function for the i-th unit evaluated at $\hat{\boldsymbol{\theta}}$, and $S(\hat{\boldsymbol{\theta}})'$ is the first numerical derivative of the score vector evaluated at $\hat{\boldsymbol{\theta}}$.

As far as model selection is concerned, the model is estimated for different values of G and D. The optimal solution is determined by selecting the model with the smallest value of the chosen information criterion, such as the Bayesian Information Criterion (BIC; [17]):

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + \nu \log(N),$$

with ν being the number of free parameters, and $\ell(\hat{\boldsymbol{\theta}})$ the log-likelihood function evaluated at $\hat{\boldsymbol{\theta}}$.

4 Conclusion

In this paper, we introduce a model-based biclustering approach designed for general variables with (conditional) distributions in the Exponential Family, observed across a sample of units. Our proposal employs a Mixture of Generalized Latent Trait Models to partition units into a specified number of components. Within each component, we use a flexible parametrization of the canonical parameter to partition variables into a given number of segments. Dependence between variables is modeled via a multidimensional, continuous latent trait,

akin to the Mixture of Latent Trait Analyzers (MLTA). In addition, covariates are incorporated into the latent layer of the model to explore their influence on component formation. Model parameter estimates are obtained using an EM algorithm, supplemented by Gauss-Hermite quadrature to approximate multidimensional integrals.

References

1. Alfò, M., Marino, M.F., Martella, F.: Biclustering multivariate discrete longitudinal data. Stat. Comput. **34**, 1–21 (2023)
2. Bock, H.: Simultaneous clustering of objects and variables. In: Tomassone, R. (eds.) Anal. des donnees es et Inform. 187–204. INRIA, Le Chesnay, France (1979)
3. Cai, J.H., Song, X.Y., Lam, K.H., Ip, H.S.: A mixture of generalized latent variable models for mixed mode and heterogeneous data. Comput. Stat. Data Anal. **55**, 2889–2907 (2011)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Series B. **39**, 1–38 (1977)
5. Failli, D., Marino, M.F., Martella, F.: An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks. In: Chelli, F.M., Ciommi, M., Ingrassia, S., Mariani, F., Recchioni, M.C. (eds.) Book of short Papers SIS 2023, Pearson (2023)
6. Failli, D., Marino, M.F., Martella, F.: Finite mixtures of latent trait analyzers with concomitant variables for bipartite networks: an analysis of COVID-19 data. Multivariate Behav. Res. **59**, 801–817 (2024)
7. Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. **24**, 569–588 (2014)
8. Good, I.: Categorization of Classification. Mathematics and Computer Science in Biology and Medicine. Her Majesty's Stationary O ce, London (1965)
9. Hartigan, J.A.: Direct clustering of a data matrix. J. Am. Stat. Assoc. **67**, 123–129 (1972)
10. Hartigan, J.A.: Clustering Algorithms. Wiley, London (1975)
11. Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Series B. **44**, 226–233 (1982)
12. Martella, F., Alfò, M.: A finite mixture approach to joint clustering of individuals and multivariate discrete outcomes. J. Stat. Comput. Simul. **87**, 2186–2206 (2017)
13. McParland, D., Gormley, I.C.: Model based clustering for mixed data: clustMD. Adv. Data Anal. Classif. **10**, 155–169 (2016)
14. Oakes, D.: Direct calculation of the information matrix via the EM. J. R. Stat. Soc. Series B. **61**, 479–482 (1999)
15. Pinheiro, J.C., Bates, D.M.: Approximations to the log-likelihood function in the nonlinear mixed-effects model. J. Comput. Graph. Stat. **4**, 12–35 (1995)
16. Royall, R.M.: Model robust confidence intervals using maximum likelihood estimators. Int. Stat. Rev. **54**, 221–226 (1986)
17. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
18. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models (1st ed.). Chapman and Hall/CRC, (2004)
19. White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. J. Econom. **48**, 817–838 (1980)



Detecting Fake News from Text: A Stagewise Methodology

Matteo Farnè[1,2(✉)] and Giulia Benelli[1,2]

[1] University of Bologna, Bologna, Italy
matteo.farne@unibo.it
[2] University of Bologna, Dipartimento di Scienze Statistiche, 40126 Bologna, Italy

Abstract. Fake news is a lasting issue in our society due to their propagation speed and their impact on public opinion. In recent years, the internet and social media have offered a perfect ground for misinformation to spread, as people can comment, elaborate and share fake news without any control. Consequently, detecting and preventing the dissemination of fake information has indeed become a pressing policy issue to tackle. For this reason, in this paper we briefly review existing fake news datasets and we present an integrated statistical methodology for fake news detection, based on text mining and classification. An application on the ISOT Fake News dataset, regarding 2016 US presidential elections, shows that ensemble methods are the most reliable in classifying fake news articles from their textual content.

Keyword: Fake News Detection, Fake News Datasets, Text Mining, Text Classification

1 Introduction

In recent years, with the advent of social media and the internet, the way information is expressed, disseminated and shared has drastically changed, with less control over the reliability of the content and the sources from which it originates. Nevertheless, information plays a fundamental role in shaping people's opinions and decisions. For this reason, it is often manipulated and distorted to control people's thinking through the creation of fake news, which is any information that is distorted and falsified with malicious or satirical intent to plagiarize others, according to the definition of [1]. The widespread dissemination of fake news can cause tangible damage to society by spreading distorted information on political, social and economic issues. It has therefore become increasingly important to develop tools and methods to detect, combat and prevent Fake News.

The term "fake news" has become extremely common in everyday language, although it represents a very complex phenomenon to define. The use of term "fake" translates as "false", "distorted" or "falsified", while the term "news" refers to all news, especially political or social ones, that are now widely spread via new mass media like the World Wide Web and social networks like Facebook

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 640–645, 2025.
https://doi.org/10.1007/978-3-031-64431-3_107

and Twitter. Even though the new digital environment has increased the speed at which news disseminate, the fake news phenomenon is still deeply rooted in a very natural and primitive human behaviour, namely the exchange of rumours and gossip [2]. Today, the same logic is used to spread information with malicious or manipulative intentions, taking advantage of the global reach of social networks, sometimes to steer people's opinions and consensus. In contrast to gossip, fake news has three characteristic features that are represented by [3]:

1. velocity, which makes it difficult to recognise and prevent fake news, as it appeals to people's emotions and therefore spreads fake information faster than real news;
2. volume, which refers to the ease with which news can be manipulated from existing articles and the global access to the internet that allows information to be replicated, acted upon and propagated;
3. veracity, which is uncertain and refers to a certain degree to which a news article can be false.

Further characteristics can be associated with fake news, such as the malicious intent that drives their authors through which fabricated and manipulated news can be identified [4] as opposed to different kinds of false news that can contain naive mistakes [5]. Moreover, textual connotations can make an article deceptive, although they are not easily recognizable. In the context of research on the linguistic cues of fake news, it is now well established that the language of liars – and those with deceptive intentions in general – is different from that of those who write with honest intent. Generally, fake news leverages certain human emotions, such as anger, fear, compassion, and trepidation [3] to elicit strong reactions in the readers and push them to propagate the news further. For this reason, fake news presents exaggerated, often inaccurate, language and negative sentiment. In [6], the authors select twelve language indicators that can be considered predictors of falsehood. They can be summarized as:

1. elusive language, with unclear or intentionally inaccurate temporal and spatial references;
2. preference for expressions and words with negative sentiment;
3. verbal, spelling and grammatical inconsistencies.

2 Fake News Datasets

Nowadays, it has become particularly easy to access articles and news through the internet and social networks, particularly through Twitter and Facebook. This fact has facilitated collecting news in large datasets, with the purpose to detect fake news. Here, we list some datasets of particular importance in this area [7]:

– LIAR [8]: a publicly available dataset containing more than 12,800 short texts collected and manually verified by "politifact.com". Statements are associated with a degree of truthfulness, as well as other manually collected information like topic, date, context, and who made the statement.

- BuzzFeedNews [9]: it is a dataset available on "github.com", which collects Facebook posts published by various media and news outlets, which cover the 2016 U.S. presidential election as their main topic. Specifically, the dataset contains articles published the week before the election and manually verified by BuzzFeed journalists.
- Kaggle FakeNews Dataset [10]: a dataset made available on "kaggle.com" that contains a dichotomous label with the "reliable" and "unreliable" modes, in addition to the title, article ID number, author, and news text. However, it is not possible to trace the sources of articles and verify their accuracy.
- CREDBANK [11]: a dataset made publicly available by researchers Mitra and Gilbert [51] that collects, as of 2015, 60 million tweets verified by thirty experts and divided into 1,049 topics.
- FakeNewsNet [12]: an archive containing two datasets, one concerning political news and the other containing gossip news. Compared to other datasets, FakeNewsNet reports textual and linguistic data along with spatiotemporal and context data.
- ISOT Fake News Dataset [13]: it consists of two datasets that collect a total of 44,898 articles related to the 2016 U.S. presidential election from various sources, including Reuters. Each dataset is related to a label, that identifies the article and can be either "Fake" or "True".
- B.S. Detector [14]: a dataset available on "kaggle.com", automatically created by the browser extension of the same name designed for the immediate detection of the veracity of News. Compared to other datasets mentioned above, the labels are not manually affixed.
- FEVER [15]: a dataset containing 185,445 sentences, some of which were altered from sentences extracted by "Wikipedia.org", which were then manually verified by assigning a label with the three classes "supported", "refuted" and "not enough info".
- Some-like-it-hoax [16]: a dataset that collects 15,500 Facebook posts and user data of those who posted the information. The posts cover a wide variety of topics, including science-related posts, which are believed to be true, and science-related fake posts or posts related to conspiracy theories.

3 Proposed Methodology

In this paper, we propose a hybrid approach to fake news detection, based on a combination of natural language processing and machine learning (see Fig. 1). As a first step, data is collected from a publicly available dataset that contains news articles labeled as either "Fake" or "True". On these articles, text pre-processing is carried out to harmonize the textual content of the news so that it becomes easier to interpret and to analyze. At the end of this phase, we obtain a harmonized corpus of documents.

As a second step, we apply a Latent Dirichlet Allocation model [17] through which we extract topics and we obtain their topic-document distributions. Contextually, a hierarchical cluster analysis is applied on extracted topic-document

distributions, in order to summarize topic information and to derive general subjects that can be later associated with the risk of the article being fake.

As a third step, extracted topic-document distributions are then manipulated through a binarization rule that permits to obtain dummy variables, which indicate the prevalence of a topic (or a cluster of topics) when the corresponding prevalence in the news article is above a pre-determined cut-off.

In the end, topic and cluster dummies become the predictors of different types of classifiers, i.e. logistic regression, classification trees and ensemble methods. By applying different classifiers, we can determine which topics or subjects are more prone to be the argument of fake news, and we can predict the probability to be fake of a given article by its textual content.

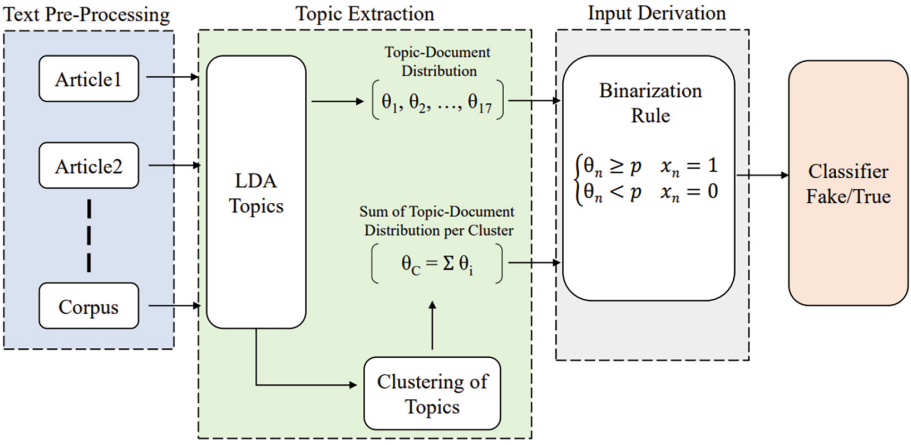


Fig. 1. Proposed statistical approach for fake news detection.

The proposed methodology has been applied to the ISOT Fake News Dataset, which is in the form of two csv files: "True.csv", containing 21417 true news items from verified and trusted sources (mainly "reuters.com"), and "Fake.csv", containing 23481 fake articles verified with "politifact.com". Table 1 contains some goodness of classification metrics for tested methods. When the predictors are topic dummies, we observe the following patterns:

- bagging and random forest present the highest accuracy, precision and recall;
- classification tree is doing worse than competitors in accuracy and precision, while it excels in recall;
- all in all, bagging presents the best F_1-score, closely followed by random forest.

When the predictors are cluster-of-topics dummies, ensemble methods still excel compared to the classification tree.

Table 1. ISOT Fake News Dataset: goodness of classification metrics for tested methods with $p = 0.1$ cut-off as binarization rule.

Logistic Regression	Accuracy	Precision	Recall	F_1-score
Topic	0.889	0.903	0.881	0.892
Cluster	0.861	0.860	0.876	0.868
Classification Tree				
Topic	0.854	0.830	0.906	0.866
Cluster	0.856	0.818	0.930	0.870
Bagging				
Topic	0.898	0.911	0.905	0.902
Cluster	0.865	0.867	0.875	0.871
Boosting				
Topic	0.895	0.905	0.892	0.889
Cluster	0.864	0.867	0.869	0.870
Random Forest				
Topic	0.898	0.911	0.905	0.901
Cluster	0.865	0.867	0.875	0.871

References

1. McGonagle, T.: "Fake news": false fears or real concerns? Neth. Q. Human Rights **35**(4), 203–209 (2017)
2. Burkhardt, J.M.: History of fake news. Libr. Technol. Rep. **53**(8), 5–9 (2017)
3. Zhang, X., Ghorbani, A.: An overview of online fake news: characterization, detection, and discussion. Inf. Process. Manage. **57**(2), 102025 (2020). https://doi.org/10.1016/j.ipm.2019.03.004
4. Tandoc, E.C., Lim, Z.W., Ling, R.: Defining "fake news". Digit. J. **6**, 137–153 (2018)
5. Ahmed, D., Hinkelmann, K., Corradini, F.: Development of fake news model using machine learning through natural language processing. arXiv preprint arXiv:2201.07489 (2022)
6. Bachenko, J., Fitzpatrick, E., Schonwetter, M.: Verification and implementation of language-based deception indicators in civil and criminal narratives. In: Coling (eds.) Proceedings of the 22nd International Conference on Computational Linguistics, pp. 1, 41–48 (2008). https://doi.org/10.3115/1599081.1599087
7. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: KNN model-based approach in classification. In: Proceedings of the OTM Confederated International Conferences, pp. 986–996. Springer, Berlin Heidelberg (2003)
8. Wang, W.: A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol.2, no. 17 (2021). https://doi.org/10.18653/v1/P17-2067.
9. Bevendorff, J., Kiesel, J., Stein B.: BuzzFeed-Webis Fake News Corpus 16 (2018). https://doi.org/10.5281/zenodo.1181813
10. Kaggle Fake News Dataset. https://www.kaggle.com/competitions/fake-news/data

11. Mitra, T., Gilbert, E.: CREDBANK: a large-scale social media corpus with associated credibility annotations. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, no. 1, pp. 258–267 (2021)
12. Shu, K., Mahudeswaran, D., Wang, S., Lee, D.: Liu, H: Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**, 171–188 (2020). https://doi.org/10.1089/big.2020.0062
13. Ahmed, H., Traore, I., Saad, S.: Detecting opinion spams and fake news using text classification. Secur. Priv. **1**(1), e9 (2018)
14. BS Detector Dataset. https://www.kaggle.com/mrisdal/fake-news
15. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: Fever: a large-scale dataset for fact extraction and verification. arXiv preprint arXiv:1803.05355 (2018)
16. Tacchini, E., Ballarin, G., Della Vedova, M., Moret, S., Alfaro, L.: Some like it hoax: automated fake news detection in social networks. arXiv preprint arXiv:1704.07506 (2017)
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2017)

Examining Socioeconomic Inequalities in Healthy Working Life Expectancy and the Association with Smoking Behaviour

Alessandro Feraldi[1(✉)] and Christian Dudel[2]

[1] MEMOTEF Department, Sapienza University of Rome, Rome, Italy
alessandro.feraldi@uniroma1.it
[2] Max Planck Institute for Demographic Research, 18057 Rostock, Germany

Abstract. Many governments worldwide are raising retirement ages due to aging populations, but this may have implications for both social and health outcomes. Health and work are influenced by socioeconomic factors. Using U.S. Health and Retirement Study data, employing a multistate lifetable approach to model individual life courses, we examine disparities in healthy working life expectancy concerning race/ethnicity, education, and smoking habits at ages 50. Findings reveal healthy working life expectancy disparities linked to socioeconomic factors and smoking behaviour, with men and non-smokers having longer healthy working life expectancy. The largest smoking differential is observed among American Indian/Alaskan Native, Asian/Pacific Islander, and individuals of other ethnicities, and Black Americans report the overall shortest healthy working life expectancy. Lower educated individuals report 1.2–1.5 years shorter healthy working life expectancy than higher educated. Addressing socioeconomic inequalities in healthy working life expectancy and promoting non-smoking habits are crucial for the sustainability of the workforce.

Keywords: Healthy Working Life Expectancy · Socioeconomic Inequalities · Smoking Behaviour

1 Introduction

The aging of the population and the increase in the old-age dependency ratio have driven numerous governments to enact or plan policies designed to raise the statutory retirement age, often entailing the elimination or reduction of early retirement benefits.

We acknowledge 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" [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. This resource was co-financed by the Next Generation EU. Progetto finanziato dal Piano Nazionale di Ripresa e Resilienza – Missione 4, "Istruzione e Ricerca" – Componente 2, "Dalla ricerca all'impresa" – Linea di investimento 1.3 – Avviso Pubblico D.D. n.341/2022 Ministero dell'Università e della Ricerca.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 646–652, 2025.
https://doi.org/10.1007/978-3-031-64431-3_108

For example, Germany, Spain, and the United Kingdom will raise their retirement age to 67 from 2023 to 2029 [5]. While these changes extend individuals' participation in the labor force, they can potentially result in higher social costs, particularly associated with adverse consequences of aging, such as unhealthy working conditions, unemployment, and reliance on disability benefits [2]. Notably, the impact of health on the individual ability to participate in work becomes particularly significant at age 50 and older [5]. This factor must be considered if governments aim to achieve and maintain sustainable extended careers.

In this context, healthy and unhealthy working life expectancy are relevant policy indicators. Healthy Working Life Expectancy (HWLE) measures the time spent working in good health, while Unhealthy Working Life Expectancy (UWLE) tracks the time spent working in poor health [3; 6]. These indicators help assess progress in extending working life and its potential impact on health when compared across countries over time. Moreover, HWLE and UWLE provide insightful summaries of individual working trajectories and they show how inequalities accumulate during the life course [3]. Trends in healthy working life expectancy exhibit substantial variations among countries. For example, it increased in England but declined in the U.S. for men [1]. Additionally, unhealthy working life expectancy increased in most countries. Possible explanations include improved treatment allowing unhealthy individuals to work longer and financial pressures forcing them to stay in the labor market.

Notably, sociodemographic factors are also linked to health and work outcomes [3; 7]. Among the potential factors of working life expectancy are the institutional setting and economic conditions and the individual health behaviours and preferences [3; 6]. For instance, research has shown that smoking and other unhealthy behaviours (e.g., lack of leisure-time physical activity, overweight, exposure to high physical work demands) increase the risk of shorter employment trajectories [6], and they might have long-lasting effects across the life course even if stopped. Nevertheless, sociodemographic and behavioural inequalities in healthy working life expectancy have received limited attention and little is known about the factors that shape this relevant measure.

In this study, the aim is to estimate healthy and unhealthy working and non-working (inactive) life expectancy at age 50, considering differences by race/ethnicity, education levels, and smoking behaviour. The study aims to shed light on the disparities in the duration of healthy working years across different behavioural and socioeconomic groups in the U.S. using a large follow-up. In doing so, we link individuals' health trajectories with their working life courses, and provide insights into life course inequalities which are crucial for policymakers aiming at extending the length of working life.

2 Data and Method

2.1 Sample

The study uses data from the RAND version of the Health and Retirement Study (HRS), an ongoing nationally representative longitudinal survey of health characteristics of U.S. men and women aged 50 and older, with oversampling of minority ethnic groups [4]. Participants have been interviewed approximately every two years from 1992 to

2020. RAND HRS version is a user-friendly longitudinal data file, which is cleaned and compiled by the RAND Corporation [2; 5].

Because the health measure of interest (i.e., limitation in five activity of daily living, more details in the subsection of measurements) was not available at the first wave of the survey, the initial sample included 38,354 individuals, interviewed between 1993/1994 (wave 2) and 2020 (wave 15) aged 50 years and older. Refreshment samples at each wave were also included. After excluding 2,318 individuals (6.0%) due to missing information on labor force, race/ethnicity, education, and smoking behaviour, the final sample included 36,036 individuals (53.2% women and 46.8% men).

2.2 Measurements

Employment is measured based on self-reported labor force state and distinguished between two states: working and inactive. Inactive respondents are those who report to be retired, out of the labor force, or unemployed. Health was measured according to limitation in five activity of daily living (ADL): bathing, eating, dressing, walking across a room, and getting in or out of bed. An individual is defined healthy (unhealthy) if he/she report none (at least one) limitation.

According to the available information on the HRS database, four race/ethnic groups are considered: Non-Hispanic White, Non-Hispanic Black, Hispanic, and Others (i.e., American Indian/Alaskan Native, Asian/Pacific Islander, or something else). Educational level is measured using the highest degree and dichotomized into higher educated (i.e. college or university degree) and lower educated (i.e. less than high-school, high-school graduate, and general equivalency degree). Finally, smoking status is dichotomized into current smoker and current non-smoker.

2.3 Analysis

The combination of two employment states and two health states allowed us to identify four living states, i.e. working healthy (state 1), working unhealthy (state 2), not working/inactive and healthy (state 3), not working/inactive and unhealthy (state 4). We used multistate Markov survival models including death as absorbing state (state 5), and consequently 16 possible transitions. Data from multiple waves were pooled into a single dataset for analysis.

The age-specific transition rates were modelled through multinomial logistic regression, using age as covariates. Educational level and smoking (time-varying) are included in the models as covariates, and the model is estimated separately for subsamples defined by gender and race/ethnicity. This implicitly interacts all covariates with gender and race/ethnicity. Transition rates are subsequently applied to a synthetic cohort in order to summarize them into duration: Total Life Expectancy (TLE); Healthy Working Life Expectancy (HWLE) - expected average number of remaining years of life in workforce and without physical limitation; Unhealthy Working Life Expectancy (UWLE) - expected average number of remaining years of life in workforce and with physical limitation; Healthy Inactive life expectancy (Inact-H) - years of life in not workforce and without physical limitation; Unhealthy Inactive life expectancy (Inact-U) - years of life in not workforce and with physical limitation.

3 Preliminary Findings

As expected, the findings of the study reveal disparities in HWLE based on gender, race/ethnicity, education, and smoking status. In Table 1 we present detailed information regarding life expectancy estimates at age 50, such as TLE, HWLE, UWLE, Inact-H, and Inact-U life expectancy. These estimates are categorized by gender, race/ethnicity, education, and smoking behaviour.

Overall, HWLE is greater in men than in women and among non-smokers compared to smokers. Additionally, approximately 95.0% of the remaining working life is expected to be without limitations in ADL. Remarkably, HWLE varies significantly across educational levels and among different racial and ethnic groups. Individuals with lower educational attainment, when compared to their higher-educated counterparts, report approximately 1.2 to 1.5 years shorter HWLEs. Overall, at age 50, non-smoking men with higher education can expect around 12.2 years of working life without limitations, whereas lower-educated men have 10.7 years. Similarly, for women, it is 11.3 years for those with higher education and 9.9 years for lower-educated women. Among smokers, there is a roughly 2-years gap for men and a 1.5-year gap for women.

Among non-smoking men, the shortest and longest HWLEs are observed among Blacks and Hispanics, respectively, with Whites and other racial/ethnic groups falling in between. For instance, among the higher educated, HWLE is 11.5 years for Black men, around 12.0 years for White and other racial/ethnic groups, and 12.7 years for Hispanic men. Among smoker men, Whites and Hispanics report the longest HWLEs, approximately 10.0 years in the higher-educated group, while it is lower among Blacks and other ethnicities, at around 9.0 years for the same educational level group. In women, HWLE at age 50 is approximately 10.5 years for non-smoking White, Black, and Hispanic women, and about 11.6 years for other racial/ethnic groups. Among White, Black, and Hispanic smoker women, HWLE ranges from 9.0 to 10.0 years, whereas it is less than 8.0 years in the other ethnic group. In both genders, the largest difference in HWLE between smokers and non-smokers, referred to as the smoking differential, is observed among the "other ethnicity" group, with a difference of 2.9 years in men, 4.0 years in higher-educated women, and 3.6 years in lower-educated women. For White, Black, and Hispanic men, the smoking differential ranges from 1.8 to 2.6 years, while in White and Black women, it is lower, at 1.4 and 1.7 years, respectively, and even lower, at 0.7 years, among Hispanic women.

Significant variations in HWLE are also noted when considering race/ethnicity, gender, and smoking behaviour in the context of education. In higher-educated White, Black, and other ethnic women, HWLE is over one year longer than in the lower-educated group. In males, educational disparities are less pronounced among Black and other ethnicity groups, with differences of 0.4 and 0.8 years, respectively. For women, smoking-related differences are slightly larger in the higher-educated Black and other ethnicity groups compared to their lower-educated counterparts, at 1.8 vs. 1.6 years (Black) and 4.0 vs. 3.6 years (other ethnicity groups), respectively. Among lower-educated men, smoking differences are slightly larger in the higher-educated group, with differences of 1.9 vs. 1.7 years in White, 2.4 vs. 2.2 years in Black, and 2.7 vs. 2.6 years in Hispanic individuals. Overall, the most significant educational gradients are observed among White men,

Table 1. Total Life Expectancy (TLE), Healthy Working Life Expectancy (HWLE), Unhealthy Working Life Expectancy (UWLE), Healthy inactive life expectancy (Inact-H), and Unhealthy inactive life expectancy (Inact-U) estimates at age 50, by gender, race/ethnicity, education, and smoking behaviour, between 1993 and 2020.

Race/Ethnicity	Education	Non-smoker					Smokers				
		TLE	HWLE	UWLE	Inact-H	Inact-U	TLE	HWLE	UWLE	Inact-H	Inact-U
Males											
Total											
High		28.1	12.2	0.4	13.2	2.3	22.4	10.1	0.4	9.5	2.4
Low		25.7	10.7	0.6	11.3	3.2	20.5	8.8	0.5	8.0	3.2
White											
High		28.6	12.3	0.4	13.7	2.3	22.6	10.4	0.4	9.5	2.3
Low		26.5	10.7	0.6	12.2	3.1	20.8	8.9	0.6	8.3	3.0
Black											
High		25.3	11.5	0.4	10.8	2.7	21.7	9.1	0.3	9.3	3.0
Low		22.5	10.1	0.4	8.9	3.2	19.3	7.9	0.3	7.5	3.5
Hispanic											
High		27.8	12.7	0.7	11.6	2.9	24.8	10.0	0.5	10.9	3.4
Low		25.6	12.2	0.8	9.1	3.6	22.6	9.6	0.5	8.4	4.1
Other											
High		26.0	12.0	0.3	11.3	2.4	21.5	9.1	0.7	9.8	1.9
Low		24.1	11.2	0.4	8.7	3.8	19.7	8.3	1.0	7.5	3.0
Females											
Total											
High		31.3	11.3	0.4	15.9	3.8	25.7	9.7	0.5	11.9	3.7
Low		28.9	9.9	0.5	13.8	4.7	23.6	8.5	0.6	10.1	4.5
White											
High		32.1	11.2	0.4	16.9	3.6	26.1	9.8	0.4	12.5	3.5
Low		29.8	10.0	0.5	15.1	4.2	24.1	8.6	0.5	11.0	4.0
Black											
High		28.1	11.1	0.6	12.2	4.2	24.1	9.4	0.6	10.0	4.1
Low		26.0	9.6	0.6	10.0	5.8	22.3	8.0	0.6	8.1	5.6
Hispanic											
High		31.4	11.5	0.8	15.2	4.0	30.4	10.7	0.8	13.6	5.3
Low		27.7	9.6	0.9	11.3	5.9	27.2	8.9	0.8	9.8	7.6
Other											
High		29.8	12.4	0.5	12.7	4.2	22.2	8.5	0.7	9.5	3.6
Low		27.9	10.8	0.4	11.6	5.1	20.5	7.2	0.6	8.5	4.2

at 1.5 years, and Hispanic women, at 1.8 years. In addition, educational differences in HWLE are consistent between smokers and non-smokers.

Disparities in HWLE are shown in Fig. 1, which displays age-specific HWLEs (in years) between ages 50 and 80, categorized by gender, race/ethnicity, and smoking behaviour, among higher and lower educated individuals, respectively. While disparities in HWLE diminish at older ages, they remain remarkable until approximately age 65.

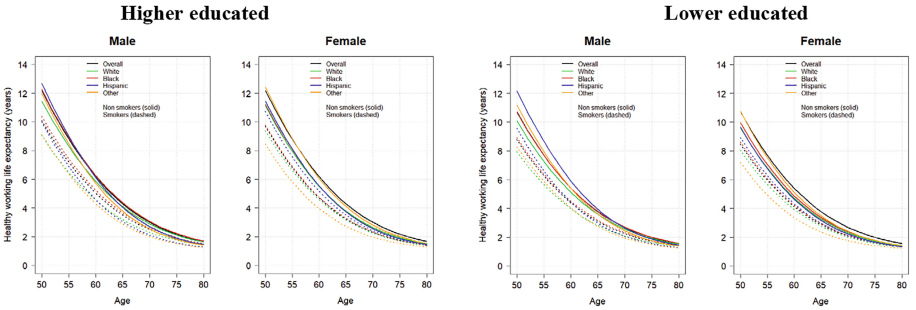


Fig. 1. Healthy Working Life Expectancy (HWLE, in years) over ages, by gender, race/ethnicity, and smoking behaviour among higher and lower educated individuals.

4 Conclusion

Our research has uncovered a significant association between smoking and a shortened span of productive and healthy working years. Smoking was found to decrease healthy working life expectancy by more than two years. Our results underscore the presence of disparities among different educational and racial/ethnic groups, with an educational gradient in healthy working life expectancy of more than one year. While individuals at the age of 50 can typically expect more than 11.0 years of healthy working life expectancy, women and Black individuals, on average, have approx. 10.0 years.

Addressing socioeconomic inequalities in healthy working life expectancy and promoting non-smoking habits are crucial for the sustainability of the workforce. These findings provide added insights given the expected increase in the statutory retirement age in numerous countries, emphasizing the urgent need to address issues in high-risk occupational sectors.

References

1. Boissonneault, M., Rios, P.: Changes in healthy and unhealthy working life expectancy over the period 2002–17: a population-based study in people aged 51–65 years in 14 OECD countries. Lancet Healthy Longevity **2**(10), e629–e638 (2021)
2. Dudel, C.: Healthy and unhealthy working life expectancy: opportunities and challenges. The Lancet Healthy Longevity **2**(10), e604–e605 (2021)
3. Hayward, M.D., Lichter, D.T.: A life cycle model of labor force inequality: extending Clogg's life table approach. Sociol. Methods Res. **26**(4), 487–510 (1998)
4. Juster, F.T., Suzman, R.: An overview of the health and retirement study. J. Hum. Resour., S7–S56 (1995)

5. Parker, M., Bucknall, M., Jagger, C., Wilkie, R.: Population-based estimates of healthy working life expectancy in England at age 50 years: analysis of data from the English Longitudinal Study of Ageing. The Lancet Public Health **5**(7), e395–e403 (2020)
6. Shiri, R., Hiilamo, A., Lallukka, T.: Indicators and determinants of the years of working life lost: a narrative review. Scand. J. Public Health **49**(6), 666–674 (2021)

Monitoring All-Cause and Cause-Specific Mortality Inequalities in Italian Regions

Alessandro Feraldi[1](✉), Marilena Pappagallo[2], Cristina Giudici[1], and Luisa Frova[2]

[1] MEMOTEF Department, Sapienza University of Rome, Rome, Italy
alessandro.feraldi@uniroma1.it
[2] Istituto Nazionale di Statistica, Rome, Italy

Abstract. Reducing socioeconomic disparities in mortality is a critical public health concern. Recognition of the importance of socioeconomic disparities in health is crucial to establish effective systems for monitoring health inequalities.

Since 2016, Istat has been tracking social inequalities in mortality using census data. Recently Istat launched a project to develop cross-sectional mortality indicators by socioeconomic conditions. Educational level is widely considered a proxy for socioeconomic status due to its associations with employment, income, and health determinants. Istat integrated data from the National Register of Causes of Death with the level of education available from the National Base Register of Individuals. The study focuses on individuals aged 30+. Standardized mortality rates and rate ratios were calculated. Between 2019 and 2020 deaths increased by 104,314 cases, with a 15% rise in the standardized mortality rate. This increase was not uniform across the country, with higher excess mortality in the Northwest. The Covid-19 impact in 2020 was significant, accounting for 10.6% of total mortality, with gender-based variations. We observed a marked gradient in mortality by education, highlighting the negative influence of disadvantaged socioeconomic status on health. From the results emerges that among females in most of the regions the educational gradient in mortality become more important when Covid-19 in considered. Findings indicate the importance of the analyses of regional disparities, accounting for age and causes of death, in order to better understand mortality inequalities in Italy. Regional variations in healthcare access and policies may amplify or alleviate health disparities, as Italy's healthcare is regionally managed.

Keywords: Cause of death · health · inequality · mortality · socioeconomic gradient

We acknowledge 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" [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. This resource was co-financed by the Next Generation EU. Progetto finanziato dal Piano Nazionale di Ripresa e Resilienza - Missione 4, "Istruzione e Ricerca" - Componente 2, "Dalla ricerca all'impresa" - Linea di investimento 1.3 - Avviso Pubblico D.D. n.341/2022 Ministero dell'Università e della Ricerca.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 653–658, 2025.
https://doi.org/10.1007/978-3-031-64431-3_109

1 Introduction

Socioeconomic inequalities in health can be defined as systematic differences in mortality rates between people of higher and lower socioeconomic status [3, 5].

Reducing socioeconomic inequalities in mortality is a public health challenge [2, 4]. The economic crisis and the most recent Covid-19 pandemic exacerbated inequalities in Italy and the gap between European countries. In addition, the importance of studies on the impact of socioeconomic inequalities on health has increased over the years. All this considered, it is relevant to have robust health inequality monitoring systems to collect relevant and essential data to produce the evidence and engage policymakers to develop effective and appropriate responses and evaluate progress [5, 7].

Since 2016 the Italian National Institute of Statistics (Istat) developed a longitudinal observation system of social inequalities in mortality (Census based). However, we needed a cross-sectional indicator system to monitor the changes in social inequalities in mortality by cause over time. Consequently, in 2023, Istat carried out a new project to develop cross-sectional indicators of mortality by socioeconomic conditions and has already made them available for the year 2019 and 2020 [8]. The level of education is generally considered as one of the best proxy of the socioeconomic status for several reasons: a lower level of education is usually associated to a worse employment status and a lower income; it is available at the individual level; it remains stable over time once a certain age is reached; and it is associated with the possible determinants of health conditions: lifestyle risk factors; propensity and ability to access preventive measures and care. Furthermore, regional variations in healthcare access and policies may amplify or alleviate health disparities, as Italy's healthcare is regionally managed [1, 6].

2 Data and Method

To measure social inequalities in health, countries should have coherent and integrated information systems of the population's socioeconomic status and health events. At this aim we integrated the National Register of Causes of Death (CoD), referring to all deaths occurred in Italy in 2019 and in 2020, with the information available (such as level of education and place of residence) in the National Base Register of Individuals (BRI) for the corresponding years [8]. The linkage performance was extremely successful (about 98.5%).

In this study, we selected deaths and the population over 30 years, as the level of education remains reasonably stable from this age onwards.

The standardized mortality rate (STD) $\times$ 10,000 (Population European Eurostat Standard, 2013) and the rate ratio (RR) between two STDs were calculated. The RR measures the excess (>1) or reduction (<1) in mortality in decedents with the lowest education compared with those with the highest level of education. The RR was calculated by sex, age class, region, and cause of death.

3 Preliminary Findings

From 2019 to 2020, the number of deaths increased by 104,314 (Table 1), and the standardized mortality rate in 2020 experienced a 15.0 percent increase compared to 2019. In particular, in the Northwest area, the excess in the mortality rate was 29.0 percent, while in the South and in the Islands, it was 7.0 and 5.0 percent, respectively.

Table 1. Deaths and population by sex. Italy, 2019 and 2020.

	2019		2020	
	Deaths	Population	Deaths	Population
Males	328,752	22,523,750	354,700	20,389,575
Females	299,659	20,422,101	378,025	22,459,542
Total	628,411	42,945,850	732,725	42,849,116

As it is well known, in 2020 Italy has been hard hit by the Covid-19 pandemic, first of all in the first wave in early 2020: the mortality due to Covid-19 was 10.6% of the overall mortality, with a relevant difference between women and men, for whom the percentage was 8.9 and 12.0 respectively.

Table 2 shows the standardized total mortality rates (per 10,000 residents) and mortality rates due to Covid-19, by sex and level of education, in Italy in 2020. Analysing the differences in mortality by the level of education, a clear gradient is found for almost all other causes of death. Less educated people had a higher level of mortality than those with higher education, showing a negative effect of disadvantaged socioeconomic status on health. Total male mortality rates (per 10,000 people) range from 26.9 (among individuals with a university degree or higher) to 65.7 (among individuals with an elementary school degree or lower). Similarly, male mortality rates due to Covid-19 range from 3.6 (university degree or higher) and 6.7 (elementary school degree or lower). In females, total mortality rates range from 16.4 and 32.6, for university degree or higher and elementary school degree or lower, respectively. Interestingly, among females, the educational gradient in mortality due to Covid-19 is more important than in males. For example, females in the lowest educational group have a mortality rate that is three times higher than the mortality rate of females in the highest educational group (2.5 and 0.8, respectively).

Rate ratios (elementary degree or less vs. university degree and more) of standardized total mortality and mortality due to Covid-19, by sex and regions are reported in Table 3. The excess mortality for less educated people is relevant also across regions. Remarkably, the total mortality rate ratios (lowest-to-highest educational group) are the lowest in Emilia-Romagna (2.1) and Umbria (1.4), for males and females respectively, whereas the highest are observed in Valle D'Aosta (3.3) among males and in Friuli Venezia Giulia among females (2.8).

Compared to the total mortality rate ratios, mortality ratios (lowest-to-highest educational group) due to Covid-19 among males are similar or lower in most of the regions.

Table 2. Standardized all-cause mortality rate and mortality rate due to Covid-19, by sex, and level of education (per 10,000 residents). Italy 2020.

	All-cause mortality			Covid-19		
	Males	Females	Total	Males	Females	Total
Elementary school or lower	65.7	32.6	46.4	6.7	2.5	4.2
Middle-school	46.7	24.2	35.9	5.0	1.8	3.4
High-school	35.9	19.7	27.7	4.2	1.2	2.6
University degree or upper	26.9	16.4	21.4	3.6	0.8	2.2
Overall	41.4	22.1	31.4	4.6	1.5	3.0

Table 3. Rate ratios (elementary degree or less vs. university degree and more) of standardized total mortality and mortality due to Covid-19, by sex and regions. Italy, 2020.

	All-cause mortality		Covid-19	
	Males	Females	Males	Females
Abruzzo	2.2	1.9	2.4	2.8
Basilicata	2.6	2.6	2.7	1.4
Calabria	2.3	1.8	0.9	1.1
Campania	2.2	1.9	1.1	2.1
Emilia Romagna	2.1	1.9	1.8	2.1
Friuli Venezia Giulia	2.6	2.8	2.7	5.0
Lazio	2.5	2.0	2.3	3.2
Liguria	2.7	1.9	2.9	3.3
Lombardia	2.3	2.1	2.1	3.3
Marche	2.2	1.9	2.8	5.0
Molise	2.8	2.4	1.5	3.9
Piemonte	2.4	1.8	2.4	1.7
Puglia	2.5	1.8	1.7	5.8
Sardegna	3.2	2.1	1.4	
Sicilia	2.4	2.0	1.8	5.0
Toscana	2.6	2.2	1.5	6.8
Trento	2.3	2.1	2.3	3.2
Umbria	2.6	1.4	0.8	
Valle D'Aosta	3.3	1.6	0.7	
Veneto	2.7	2.0	3.2	9.6
Italy	2.4	2.0	1.9	3.0

Conversely, among females, mortality rate ratios due to Covid-19 are higher than the total mortality ratios in several regions, especially in Veneto (9.6), Toscana (6.8), and Puglia (5.8).

The heat maps of the Italian regions according to the rate ratios (elementary degree or less vs. university degree and more) of standardized total mortality and mortality due to Covid-19, by sex are reported in Fig. 1.

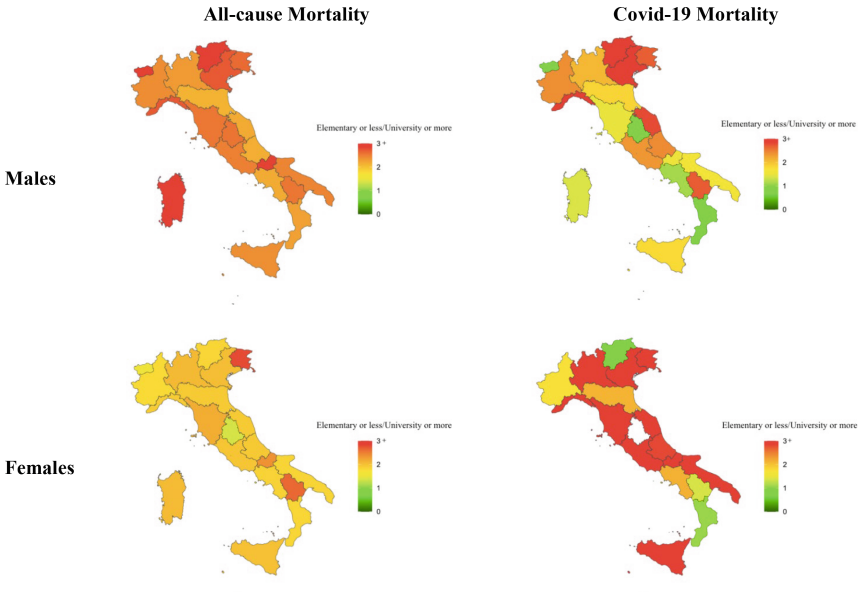


Fig. 1. Rate ratios (elementary degree or less vs. university degree and more) of standardized total mortality and mortality due to Covid-19, by sex and regions. Italy, 2020.

From the figure emerges that, among females, in most of the regions the educational gradient in mortality becomes more important when Covid-19 mortality is considered. Indeed, when considering all-cause mortality, most regions exhibit a mortality rate among females that is approximately twice as high among those with the lowest education compared to the highest education (regions colored in yellow). However, in several of these regions this ratio increases to 3 or more when examining Covid-19 mortality. The pattern differs among males, with most Italian regions showing a reduction in educational disparities in Covid-19 mortality compared to total mortality. Notably, in regions such as Valle D'Aosta, Umbria, and Calabria, Covid-19 mortality rates are higher among males with university degree or more than males with elementary degree or lower, with rate ratios of 0.7, 0.8, and 0.9, respectively (see Table 3).

The preliminary results indicate that the analysis of causes of death by age groups, sex, and regions, could better explain inequalities in mortality variations, especially over the course of the pandemic. Results suggest that regional differences in the availability or access to healthcare may exacerbate or mitigate health inequalities, especially in the context of Italy, where health policies are managed at the regional level.

In order to analyse socioeconomic inequalities by regions, further analysis using more complex socioeconomic inequality indices will be carried out.

References

1. Costa, G., Bassi, M., Gensini, G.F., Marra, M., Nicelli, A.L., Zengarini, N.: L'equità nella salute in Italia. Secondo rapporto sulle disuguaglianze sociali in sanità. Franco Angeli Editore, Milano (2014)
2. Mackenbach, J.: Health inequalities: Europe in profile. Report for 'Tackling Health Inequalities Governing for Health'. COI London (2006)
3. Mackenbach, J.P., Kunst, A.E.: Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from Europe. Soc Sci Med **44**(6), 757–771 (1997)
4. Marmot, M.: The Health Gap: The Challenge of an Unequal World. Bloomsbury Press (2015)
5. Murtin, F., Mackenbach, J.P., Jasilionis, D., Mira, D.M.: Educational inequalities in longevity in 18 OECD countries. J. Demographic Econ. **88**, 1–29 (2022)
6. Petrelli, A., Frova, L.: Italian atlas of mortality inequalities by education level. Epidemiologia e Prevenzione, anno **43**(1) (2019)
7. Renard, F., Devleesschauwer, B., Speybroeck, N., Deboosere, P.: Monitoring health inequalities when the socio-economic composition changes: are the slope and relative indices of inequality appropriate? Results Simul. Study BMC Public Health **19**, 662 (2019)
8. Tavole di Dati: Disuguaglianze Nella Mortalità Per Causa in Italia Secondo Caratteristiche Demografiche, Sociali E Territoriali - Anno (2020)

A Nonparametric Test for Fixed Effects in Crossed Random Effect Models

Federico Ferraccioli[1(✉)], Livio Finos[1], and Angela Andreella[2]

[1] Università degli Studi di Padova Dipartimento di Scienze Statistiche, Padova, Italy
{federico.ferraccioli,livio.finos}@unipd.it
[2] Università Ca' Foscari Venezia, Dipartimento di Economia, Venezia, Italy
angela.andreella@unive.it

Abstract. Mixed-effects models are powerful tools for analyzing data from experimental designs involving both subjects and items as random effects. In fields such as psychology, neuroscience, and digital markets, researchers commonly use mixed-effects models to analyze data from experiments or customer ratings. In this context, we propose a nonparametric approach based on the sign-flip score test. The method does not require distributional assumptions on the error components and remains valid even with low sample sizes and unbalanced data. The performance of the test is shown through the experimental results.

Keywords: Nonparametric test · mixed effects · crossed designs · sign-flip score

1 Introduction

Mixed effects models represent a powerful tool for analyzing data from experimental designs where both subjects and items are considered as random effects. These models are particularly useful in situations where traditional methods may not adequately account for the nested or crossed structure of the data.

In the fields of psychology and neuroscience, e.g., [2,12], researchers often use mixed-effects models to analyze data from experiments in which subjects are presented with various stimuli, such as images, sounds, or videos. Each subject's response to these stimuli can be considered as a random effect, as can the specific stimuli themselves. This allows researchers to account for the variability between subjects and between stimuli when estimating the overall effects of experimental manipulations.

Similarly, in digital markets or streaming services [5,6], mixed effects models can be used to analyze data on customer ratings of different items or content. By treating both customers and items as random effects, these models can provide insight into how individual preferences and characteristics influence ratings while also accounting for the inherent variability between items.

The crossed random effect model we consider has the form

$$y_{ij} = x_{ij}^{\top}\boldsymbol{\beta} + a_i + b_j + \varepsilon_{ij}, \quad 1 \leq i \leq R, 1 \leq j \leq C,$$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 659–664, 2025.
https://doi.org/10.1007/978-3-031-64431-3_110

where a_i and b_j are the random effects and ε_{ij} represents the error term. The fixed effect $\boldsymbol{\beta} \in \mathbb{R}^p$ is associated with the covariates $x_{ij} \in \mathbb{R}^p$. Similar to [5], we assume that $a_i \sim (0, \sigma_A^2)$, $b_j \sim (0, \sigma_B^2)$, and $\varepsilon_{ij} \sim (0, \sigma_E^2)$ are all independent.

In such contexts, where data involve the rating of R people for C items, it is common to encounter highly unbalanced data with an irregular observational pattern, where only a fraction of the covariate-response pairs are observed, i.e., $N \ll R \times C$ where N is the total number of observations. In this framework, the estimation problem for these methods has been extensively explored in the recent literature, both from Bayesian [5,9], and from frequentist perspectives [6]. In the following, we let $Z_{ij} \in \{0, 1\}$ take the value of 1 if (x_{ij}, Y_{ij}) is observed and 0 otherwise, for $i = 1, \ldots, R$ and $j = 1, \ldots, C$. Throughout the manuscript, we assume Z_{ij} to be non-random.

Our focus here lies in addressing the challenge of establishing an appropriate testing procedure for the fixed component within such models. Specifically, we propose a nonparametric approach based on the concept of eigen sign-flip [3]. A major advantage of the proposed nonparametric approach over standard parametric tests lies in the fact that it does not require distributional assumptions on the error components, thus remaining valid even in instances of low sample size and unbalanced data scenarios. Nonparametric proposals in this context (see, for example, [4,8]) are limited to simpler designs, i.e., assuming a single factor or no crossing.

The work is organized as follows. In the next Section, we outline the methodology, while in Sect. 3, we show the performances of the proposed methodology via simulations. Section 4 is devoted to a discussion and possible extensions.

2 Methodology

2.1 Definition of the Problem

In the following, we will denote by $Y \in \mathbb{R}^N$ the vector containing all the stacked values Y_{ij}, and $X \in \mathbb{R}^{N \times p}$, where p is the number of covariates. Using this notation, We can rewrite the crossed effects model as

$$Y = X\boldsymbol{\beta} + Z_A\mathbf{a} + Z_B\mathbf{b} + \boldsymbol{\varepsilon},$$

where $Z_A \in \{0, 1\}^{N \times R}$ and $Z_B \in \{0, 1\}^{N \times C}$ denotes the observed pairs for the two effects. Note that the i-th column of Z_A has ones for all the N observations that come from row i and zeros elsewhere. The definition of Z_B is analogous (see Fig. 1).

It is well known (see, for instance, [11]) that the estimate $\hat{\beta}$ can be obtained via generalized least squares (GLS) by solving the penalized least-square problem

$$\min_{\boldsymbol{\beta},\mathbf{a},\mathbf{b}} ||Y - X\boldsymbol{\beta} - Z_A\mathbf{a} - Z_B\mathbf{b}||^2 + \lambda_A||\mathbf{a}||^2 + \lambda_B||\mathbf{b}||^2. \quad (1)$$

The GLS estimate $\hat{\beta}$ in the case of crossed random effects can be efficiently estimated if all $R \times C$ are available. In the case of $N \ll R \times C$, several recent works have addressed the problem (for example, [6,9]).

2.2 Test Procedure

In this framework, a natural question is whether the covariates X have an effect on the variable of interest. We are thus interested in the system of hypotheses

$$\mathrm{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad \text{versus} \qquad \mathrm{H}_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

The standard parametric Wald type test may have poor performances in small sample scenarios and when the design of the crossed effect is highly unbalanced due to the overestimation of the variance of the test statistic induced by the penalization. In the following, we propose a nonparametric approach similar to the eigen-sign flip procedure in [3].

Let $\mathbf{g}^\top = (\mathbf{a}^\top, \mathbf{b}^\top)$, $P_\lambda = \mathrm{diag}(\lambda_A \mathbb{I}_A, \lambda_B \mathbb{I}_B)$ and $Z_G = [Z_A | Z_B]$ ($R + C$ columns). The minimization problem in (1) can be rewritten as

$$\min_{\beta,\mathbf{g}} ||Y - X\boldsymbol{\beta} - Z_G\mathbf{g}||^2 + \mathbf{g}^\top P_\lambda \mathbf{g}. \tag{2}$$

Note that the values of the two smoothing parameters (λ_A, λ_B) are strictly related to the variance components of the random effects, more specifically $\lambda_A = \sigma_E^2/\sigma_A^2$ and $\lambda_B = \sigma_E^2/\sigma_B^2$. The solution to (2) is uniquely determined by the normal equations

$$X^\top(Y - X\boldsymbol{\beta} - Z_G\mathbf{g}) = 0, \tag{3}$$

$$(Z_G^\top Z_G + P_\lambda)\mathbf{g} - Z_G^\top(Y - X\mathbf{g}) = 0. \tag{4}$$

Similar to [3], under the null hypothesis the score for the covariates from (3) is

$$X^\top(Y - X\boldsymbol{\beta}_0 - Z_G\mathbf{g}).$$

Since $\mathbf{g}$ is unknown, we can plug-in the expression for $\hat{g}$ obtained from (4) as

$$\begin{aligned} T &= X^\top(Y - Z_G\hat{\boldsymbol{g}} - X\boldsymbol{\beta}_0) \\ &= X^\top(Y - Z_G(Z_G^\top Z_G + P_\lambda)^{-1}Z_G^\top(Y - X\boldsymbol{\beta}_0) - X\boldsymbol{\beta}_0) \\ &= X^\top \Lambda(Y - X\boldsymbol{\beta}_0) = X^\top \Lambda \mathbf{r}, \end{aligned} \tag{5}$$

with $\mathbf{r} = Y - X\boldsymbol{\beta}_0$ and $\Lambda = I - Z_G(Z_G^\top Z_G + P_\lambda)^{-1}Z_G^\top$. Assuming we have a reasonable estimate of the quantities of interest, i.e., $(\hat{\beta}, \hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\sigma}_E^2)$, we can now define the test procedure.

Definition 1. *Let consider the singular value decomposition $\Lambda = VDV^\top$. Set $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_N)$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ is a random vector uniformly distributed in $\{-1, 1\}^N$. Let us also define the N-dimensional vectors $\tilde{X} = D^{1/2}V^\top X$ and $\tilde{\mathbf{r}} = D^{1/2}V^\top\mathbf{r} = D^{1/2}V^\top(Y - X\boldsymbol{\beta}_0)$. The eigen sign-flip statistics is defined as*

$$\begin{aligned} T_\Pi &= X^\top V D^{1/2} \Pi D^{1/2} V^\top (Y - X\boldsymbol{\beta}_0) \\ &= \tilde{X}^\top \Pi \tilde{\mathbf{r}}. \end{aligned} \tag{6}$$

Note that the observed statistic $T = T_I$, which corresponds to the case where $\pi_i = 1 \forall i$, is included in the set of possible sign-flips in order to get an exact control α . As standard in permutational approaches, the component-wise p-values are thus computed as the rank of T_I with respect to a sample of M sign-flips $\boldsymbol{\pi}$, divided by M [10].

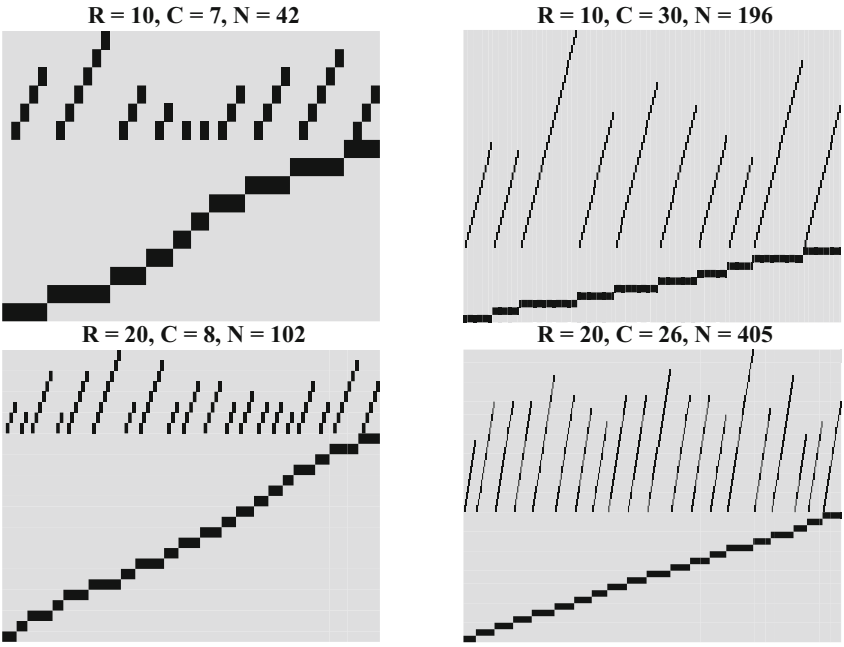


Fig. 1. Graphical representation of Z_G in the four settings. The black cells correspond to the ones, while the grey cells correspond to the zeros.

3 Experimental Results

In this Section, we present some preliminary results regarding the performances of the proposed test in different scenarios. We generate the data from the model presented in Sect. 2, with $\sigma_A = 3$, $\sigma_B = 2$ and $\sigma_E = 1$. We define four unbalanced settings, characterized by different combinations of R and C (see Fig. 1). Note that R is fixed while the number of items C is sampled from a Poisson distribution, with mean 5 (C = {7, 8}) and 20 (C = {30, 26}). Moreover, we define a single covariate X in two cases :

- Case A: x is an i.i.d. sample from a standard Gaussian;
- Case B: x is a sample from a multivariate Gaussian with mean 0 and covariance structure proportional to $Z_G Z_G^\top$.

Note that Case B should represent a more challenging case due to possible dependencies between the fixed and the random effects. We compute the control of the Type 1 error ($\beta = 0$) and the power ($\beta = \{1, 2\}$) for all the cases with 1000 random sign-flips. The results, presented in Table 1, are averaged over 200 repetitions.

Table 1. Empirical results (significance level $\alpha = 0.05$).

	R = 10, C = 7, N = 42		R = 10, C = 30, N = 196	
	Case A	Case B	Case A	Case B
$\beta = 0$	0.065	0.045	0.060	0.060
$\beta = 1$	0.675	0.255	1.000	0.935
$\beta = 2$	0.755	0.460	1.000	0.985
	R = 20, C = 8, N = 102		R = 20, C = 26, N = 405	
	Case A	Case B	Case A	Case B
$\beta = 0$	0.030	0.045	0.055	0.040
$\beta = 1$	0.700	0.260	1.000	0.895
$\beta = 2$	0.890	0.485	1.000	1.000

The results in Table 1 are promising. In both low and high sample size settings and for both cases, the test maintains the control of the Type-I error (although slightly conservative at times) and a reasonable power. As expected, the results for Case A are more convincing, especially in the low sample size setting. The particular structure of the covariance in Case B influences the estimation of the variance, thus reducing the power of the test.

4 Discussion

We proposed a nonparametric test for fixed effects in crossed random effects models. The method demonstrates promising results in different scenarios, namely unbalanced designs and low sample sizes. One of the key strengths of the proposed test lies in its ability to maintain control of Type I error rates while still exhibiting reasonable power. However, it shows some potential limitations stemming from the estimation of random effects, which could potentially induce biases and reduce power in certain situations.

To improve the efficacy of our test, future investigations should thoroughly explore the impact of design matrix structure and covariates on finite sample properties. This deeper understanding could lead to refinements that mitigate biases and improve the overall performance of the test, especially in low sample sizes. Further examinations will also involve comparing the efficacy of the proposed method with that of other state-of-the-art competitors to determine its comparative advantages and limitations.

Furthermore, the formulation of our test lends itself to extension within generalized linear models (see, for instance, [1,7]) and the multivariate framework ([12]). These avenues of research hold substantial promise and could emerge as robust alternatives to conventional parametric approaches.

References

1. Andreella, A., Goeman, J., Hemerik, J., Finos, L.: Robust Inference for Generalized Linear Mixed Models: An Approach Based on Score Sign Flipping. arXiv preprint arXiv:2401.17993 (2024)
2. DeBruine, L.M., Barr, D.J.: Understanding mixed-effects models through data simulation. Adv. Methods Pract. Psychol. Sci. **4**(1), 2515245920965119 (2021)
3. Ferraccioli, F., Sangalli, L. M., Finos, L.: Nonparametric tests for semiparametric regression models. TEST 1–25 (2023)
4. Finos, L., Basso, D.: Permutation tests for between-unit fixed effects in multivariate generalized linear mixed models. Stat. Comput. **24**, 941–952 (2014)
5. Gao, K., Owen, A.B.: Estimation and inference for very large linear mixed effects models. Stat. Sin. **30**(4), 1741–1771 (2020)
6. Ghosh, S., Hastie, T., Owen, A.B.: Backfitting for large scale crossed random effects regressions. Ann. Stat. **50**(1), 560–583 (2022)
7. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. R. Stat. Soc. Ser. B Stat Methodol. **82**(3), 841–864 (2020)
8. Kherad-Pajouh, S., Renaud, O.: A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. Stat. Pap. **56**(4), 947–967 (2015)
9. Papaspiliopoulos, O., Roberts, G.O., Zanella, G.: Scalable inference for crossed random effects models. Biometrika **107**(1), 25–40 (2020)
10. Pesarin, F.: Multivariate Permutation Tests: With Applications in Biostatistics, vol. 240. Wiley, Chichester (2001)
11. Robinson, G.K.: That BLUP is a good thing: the estimation of random effects. Stat. Sci. 15-32 (1991)
12. Visalli, A., Montefinese, M., Viviani, G., Finos, L., Vallesi, A., Ambrosini, E.: lmeEEG: mass linear mixed-effects modeling of EEG data with crossed random effects. J. Neurosci. Methods **401**, 109991 (2024)



A Bayesian Cure Model for Earthquake Parameter Estimation Using Crowdsourced Smartphone Data

Francesco Finazzi[1(✉)], Luca Aiello[2], and Lucia Paci[3]

[1] Department of Economics, Università degli Studi di Bergamo, Bergamo, Italy
francesco.finazzi@unibg.it
[2] Department of Economics, Management and Statistics, Università degli Studi di Milano-Bicocca, Milan, Italy
[3] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

Abstract. Earthquake Early Warning Systems (EEWS) are critical tools for regions prone to seismic activity. However, their widespread adoption is hampered by the high cost of traditional systems, particularly in low-income areas. Recently, researchers have proposed low-cost alternatives, such as smartphone-based EEWSs, despite the reliability challenges of smartphones. This work presents a statistical methodology for estimating key earthquake parameters using smartphone data. Borrowing from survival data analysis, a Bayesian cure model is proposed that treats smartphones as patients in a clinical trial, with earthquake detection as the censoring event. Incorporating spatial and temporal data, a mixture of parametric densities is developed to represent detectable earthquake waves. The model is fitted using an adaptive Markov chain Monte Carlo algorithm. A real-world case study demonstrates the robustness of the model and provides insights into smartphone-based earthquake monitoring.

Keywords: Earthquake early warning · survival analysis · mixture modelling · citizen science

1 Introduction

Earthquake early warning systems (EEWSs) [1] are employed in seismic-prone regions to mitigate earthquake risk. These systems are based on dense networks of instruments for real-time earthquake detection, halting critical processes, and issuing alerts to those facing strong ground shaking. However, the high cost of traditional EEWS, typically in the millions of euros, has hindered their widespread adoption, especially in low-income seismic regions.

In recent decades, researchers have proposed low-cost alternatives [2], with the advent of smartphones revolutionising the accessibility of such solutions. The

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 665–670, 2025.
https://doi.org/10.1007/978-3-031-64431-3_111

Earthquake Network (EQN) citizen science initiative [3] pioneered smartphone-based EEWS, delivering real-time alerts to users' smartphones. Since 2013, more than 19 million people have participated in the in the initiative, and the system has contributed to nearly 6,000 alerts being sent. Although smartphones are less reliable due to limitations of their accelerometers and anthropic noise, specific statistical methods have been developed for real-time earthquake detection [4] and performance assessment of smartphone-based networks [5].

EQN works through a widespread smartphone network across multiple platforms including Android, iOS and Huawei. Users install the EQN app, which uses smartphone accelerometers for seismic monitoring during periods when the devices are charging, creating a dynamic monitoring network. The data collected from these smartphones is stored on a server and consists of two lists. The list of active smartphone (namely smartphones that are known to be monitoring at the earthquake detection time but that did not trigger) and the list of triggering smartphones (smartphones that detected an acceleration).Active smartphones are described by their spatial coordinates, while triggering smartphones are described by their coordinates and triggering times. Both lists cover an area of 300 km radius centred on the EQN detection location and a period from two minutes before the EQN detection time to the detection time.

This work extends previous research by developing a statistical methodology to estimate key earthquake parameters (epicentre, depth and origin time) while addressing smartphone-specific challenges. We use survival data analysis, comparing smartphones to patients in a clinical trial, with earthquake detection as the censoring event. Drawing on population-based clinical studies, we propose a Bayesian cure model to estimate earthquake parameters, incorporating spatial and temporal information from triggering and spatial information from non-triggering smartphones. The model assumes a mixture of parametric densities for the mortality function, representing the earthquake waves detectable by smartphones. Posterior inference is performed using a tempered adaptive Markov chain Monte Carlo (MCMC) algorithm. We use a real case study to illustrate the soundness of our modelling approach and its enhancement of the EQN system.

2 Model

Let T be the triggering time of the smartphone and T^* be the censoring time, which corresponds to the moment when the earthquake is detected by EQN. Consequently, we define $Y = \min(T, T^*)$ and $\Delta = I(T < T^*)$ the censoring indicator.

In relative survival models [6,7], the overall (all-cause) survival function $S(t) = P(Y > t)$ is expressed as the product of the expected survival function $S_0(t)$ and the relative survival function $R(t)$, as

$$S(t) = S_0(t)R(t). \tag{1}$$

Hence, the overall hazard function $h(t)$ is obtained as the sum of two components: the expected hazard function $h_0(t)$ and the excess hazard function $\lambda(t)$, that is

$$h(t) = h_0(t) + \lambda(t). \tag{2}$$

The premise underlying Eqs. (1) and (2) is the assumption of independence between the triggering events related to the earthquake and those due to other causes, implying the existence of independent competing risks [8]. Equation (2) assumes that both $S_0(t)$ and $h_0(t)$ are known.

Traditionally, as time approaches infinity, the survival curve tends to 0, denoted by $S(\infty) = 0$. However, this assumption is not realistic in our context, as a non-negligible fraction of smartphones will not trigger or will not be able to send the triggering information to the EQN server infrastructure, usually due to technical problems of the smartphone or the Internet connection. Therefore, we adopt cure models, a class of models designed for censored survival data, where some individuals will not experience the event of interest regardless of the length of follow-up. These individuals who are not expected to experience the event are often referred to as *cured* subjects. See [9] for further details.

Given that some smartphones may never trigger during an earthquake event, it is reasonable to adopt a mixture model where the population consists of two distinct groups: (i) a fraction π representing cured cases (smartphones that will never trigger due to the earthquake), and (ii) a fraction $(1 - \pi)$ representing uncured cases (smartphones that may eventually trigger due to the earthquake). Let W denote the cure indicator, where $W = 1$ if the smartphone is cured, with $P(W = 1) = \pi$. Let $S_Q(t) = P(T > \text{ t} \mid W = 0)$ be the survival function of the uncured population; hence the following survival function defines the mixed cure model:

$$S(t) = S_0(t)\{\pi + (1 - \pi)S_Q(t)\}, \tag{3}$$

where the survival function with respect to the π component is one, since a cured smartphone will never trigger due to the earthquake. The corresponding hazard rate can be expressed as the sum of the background trigger rate and the earthquake-related trigger rate, i.e.

$$h(t) = h_0(t) + \frac{(1 - \pi)f_Q(t)}{\pi + (1 - \pi)S_Q(t)}, \tag{4}$$

where $f_Q(t)$ is the probability density function associated with $S_Q(t)$. Hence, the excess hazard function $\lambda(t)$ has an asymptote at zero.

Earthquakes are defined by their primary (P) and secondary (S) waves, with the P wave typically travelling faster than the S wave and generally causing mild ground shaking. During an earthquake event, a monitoring smartphone may detect either the P or S wave, depending on several factors such as the magnitude of the earthquake, the distance between the smartphone and the epicentre, and the location of the smartphone. To account for this, $f_Q(t)$ in Eq. (4) is represented as a mixture of two densities corresponding to the P and S waves. Specifically, $f_Q(t) = \alpha f_P(t) + (1 - \alpha)f_S(t)$, where $f_P(t)$ and $f_S(t)$ are the densities of normal distributions with a common variance of τ^2 and means

$$\begin{aligned} \mu_P(\boldsymbol{\theta}, \mathbf{z}_i) &= \frac{dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)}{v_P} + t_0 + \mu \\ \mu_S(\boldsymbol{\theta}, \mathbf{z}_i) &= \frac{dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)}{v_S} + t_0 + \mu, \end{aligned} \tag{5}$$

where $\boldsymbol{\theta} = (\mathbf{z}_0, d_0, t_0)$, $\mathbf{z}_0 = (lon_0, lat_0)$ denotes the epicentre, d_0 the depth, and t_0 the origin time. The location of each smartphone is denoted by $\mathbf{z}_i = (lon_i, lat_i)$, and v_P and v_S represent the wave velocities. The parameters μ and τ^2 are chosen to encompass 99% of the probability within a window of 3.5 seconds, which is the maximum delay allowed for a triggering smartphone to transmit the signal to the server. Additionally, $dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)$ denotes the Euclidean distance between smartphone i and the hypocentre $(\mathbf{z}_0, d_0)$.

Hence, the survival function becomes

$$S(t) = S_0(t)\{\pi + (1-\pi)(\alpha S_P(t) + (1-\alpha)S_S(t))\}, \tag{6}$$

with hazard function

$$h(t) = h_0(t) + \frac{(1-\pi)(\alpha f_P(t) + (1-\alpha)f_S(t))}{\pi + (1-\pi)(\alpha S_P(t) + (1-\alpha)S_S(t))}. \tag{7}$$

This model resembles a three-component mixture, similar to the approach in [7]: (i) a fraction π representing smartphones unaffected by the earthquake and never triggered; (ii) a fraction $(1-\pi)\alpha$ of smartphones detecting the P wave; and (iii) the remaining fraction $(1-\pi)(1-\alpha)$ of smartphones detecting the S wave.

Considering n active smartphones in the study area throughout the observation period, let y_i, $i = 1, \ldots, n$, denote the minimum time between the signal sent by smartphone i and the EQN detection time t^*. If y_i is uncensored ($\delta_i = 1$), the i-th smartphone contributes $f(y_i|\boldsymbol{\theta}, \mathbf{z}_i) = h(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ to the likelihood. If y_i is censored ($\delta_i = 0$), the i-th smartphone contributes $S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ to the likelihood, namely

$$\mathcal{L}(\boldsymbol{\theta}, \pi, \alpha; \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) = \prod_{i=1}^{n} h(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)^{\delta_i} S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i), \tag{8}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\boldsymbol{\delta} = (\delta_1 \ldots, \delta_n)$ and $h(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ and $S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ are as in (6) and (7).

The Bayesian hierarchy of the model is completed by specifying prior distributions for all parameters. For epicentre coordinates $\mathbf{z}_0$ we use a bivariate normal distribution centred on $\mathbf{z}^*$, the EQN detection location, with a unit diagonal covariance matrix. The depth d_0 follows a uniform distribution from 0 to 100 km, while the origin time t_0 is modelled with an exponential distribution with rate $\lambda = 1/20$ seconds^{-1}. The wave mixing weight α is governed by a beta distribution with parameters $a = 1/2$ and $b = 1/2$, assuming a priori that only one wave is detected. The cure fraction π is uniformly distributed between zero and one. This model, called Survival EarthQuake Model (SEQM), integrates these prior assumptions.

3 Analsysis of Chilean Earthquake Data

In this work, we analyse data from an earthquake that occurred in northern Chile on 29 November 2019. The parameters of the earthquake, obtained from

the European-Mediterranean Seismological Centre (EMSC; https://www.emsc-csem.org), are as follows: latitude -21.34, longitude -68.94, depth 121 km, and time of origin 11:06:05.

We fit the SEQM outlined in Sect. 2 using an adaptive MCMC algorithm to estimate the earthquake parameters. The results are summarised in Table 1, which shows the highest posterior mode and associated 95% Highest Posterior Density Regions (HPDRs) of all model parameters. The table also shows the EQN estimates of the earthquake parameters. While EQN only provides point estimates, SEQM provides uncertainty quantification through the HPDRs. The left panel of Figs. 1 shows a summary map showing the EMSC epicentre and its estimates provided by EQN and SEQM, with the latter computed as the highest posterior mode. The right panel shows the posterior distribution of the epicentre and its estimated density.

Table 1. EQN and SEQM earthquake parameter estimates. For SEQM, also the 95% HPDR is provided.

EQN			**SEQM**					
lat^*	lon^*	t^*	lat_0	lon_0	d_0	t_0	α	π
−20.25	−70.12	11 : 06 : 37	−21.23	−69.04	89.45	11 : 05 : 53	0.13	0.87
			[−21.30, −21.20]	[−69.14, −68.93]	[65.24, 99.53]	[11 : 05 : 52, 11 : 05 : 55]	[0.01, 0.48]	[0.66, 0.94]

Notably, the marginal posterior distribution of the epicentre largely encapsulates the EMSC values, as shown in Table 1. In particular, the true longitude falls within the highest posterior region, as visually highlighted in Fig. 1, while the latitude approaches inclusion.

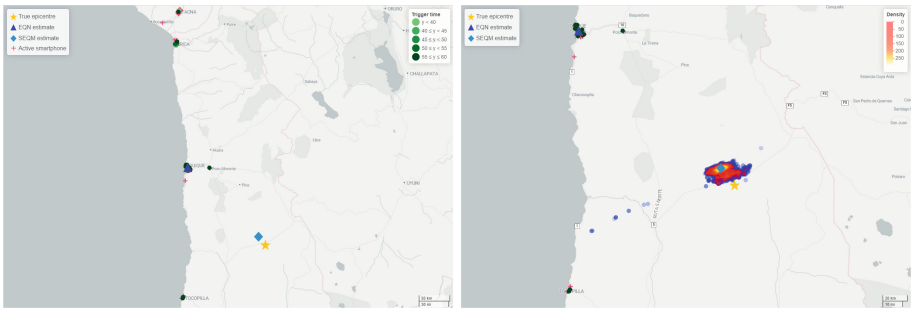


Fig. 1. Summary map (left panel) and posterior density of the epicentre location (right panel). The zero of the triggering times is the earthquake origin time.

For EQN, the epicentre error is reported to be 172.71 km, with an origin time error of 32.00 s. In contrast, SEQM shows a smaller epicentre error of 16.86 km, along with a depth error of 31.55 km and an origin time error of 16.82 s. This

comparison highlights the differences in accuracy between the two systems, with SEQM generally showing lower error values.

4 Discussion

Inspired by survival analysis, this approach accounts for uncertainties in smartphone data and uses information from both triggering and non-triggering smartphones to improve parameter estimation, especially in scenarios with limited triggers. Application to a real dataset demonstrates improved estimates of epicentre and origin time. Future directions include incorporating random variables for seismic wave velocities, implementing dependent data priors, and optimising computational efficiency for real-time use.

References

1. Gasparini P. Earthquake Early Warning Systems. Springer (2007)
2. Clayton, R., et al.: Community seismic network. Ann. Geophys. **54**(6), 738–747 (2011)
3. Finazzi, F.: The earthquake network project: a platform for earthquake early warning, rapid impact assessment, and search and rescue. Front. Earth Sci. **8**, 243 (2020)
4. Finazzi, F., Fassò, A.: A statistical approach to crowdsourced smartphone-based earthquake early warning systems. Stoch. Environ. Res. Risk Assess. **31**(7), 1649–1658 (2017)
5. Finazzi, F., Bondár, I., Bossu, R., Steed, R.: A probabilistic framework for modeling the detection capability of smartphone networks in earthquake early warning. Seismol. Res. Lett. **93**(6), 3291–3307 (2022)
6. De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., Verdecchia, A.: Mixture models for cancer survival analysis: application to population-based data with covariates. Stat. Med. **18**(4), 441–454 (1999)
7. Lambert, P.C., Dickman, P.W., Weston, C.L., Thompson, J.R.: Estimating the cure fraction in population-based cancer studies by using finite mixture models. J. R. Stat. Soc. Ser. C Appl. Stat. **59**(1), 35–55 (2010)
8. Gamel, J.W., Vogel, R.L.: Non-parametric comparison of relative versus cause-specific survival in surveillance, epidemiology and end results (SEER) programme breast cancer patients. Stat. Methods Med. Res. **10**(5), 339–352 (2001)
9. Peng, Y., Taylor, J.M.G.: Cure models. Handb. Surv. Anal. **34**, 113–134 (2014)

Analysis of Mastodon's Italian Messages: Networks, Topics, LLM and Machine Learning

Livio Finos[1] and Andrea Sciandra[1,2](✉)

[1] Department of Statistical Sciences, University of Padova, Padua, Italy
{livio.finos,andrea.sciandra}@unipd.it
[2] Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Padua, Italy

Abstract. This study examines Italian-language content on the social media platform Mastodon, employing some typical tools for textual analysis. A set of posts (known as 'toots' on Mastodon) related to the hashtag #intelligenzaartificiale from the past two years were collected. Co-occurrence networks between hashtags and cross-author referrals within the same toots were analyzed. Next, structural topic modeling was utilized to identify four topics and their most important keywords. A comparison was then made between the predictive ability of a set of words and that of a transformer using a classifier, resulting in similar findings. Furthermore, a SHAP analysis was conducted to demonstrate the impact of individual words on the classification model, providing an explanation of the contribution of individual features.

Keywords: Mastodon · topic modelling · transformers · classification

1 Introduction

The emergence of Decentralized Online Social Networks (DOSNs) represents a departure from the centralized platforms that dominate the social media landscape. Unlike their centralized counterparts, which are often controlled by single owners driven by marketing strategies, DOSNs prioritize user experience, transparency, and privacy (La Cava et al., 2022). Decentralized alternatives to mainstream social media giants are provided by platforms such as Mastodon, Diaspora, and Hubzilla. These platforms break down their services into independent servers (instances) and allow cross-instance interaction through federation (the Fediverse). Mastodon is a decentralized social network within the Fediverse that is widely recognized. It enables users to create communities based on shared interests and communicate easily between different instances, promoting a more diverse and decentralized online environment (He et al., 2023). Furthermore, Mastodon provides a platform for users to express themselves freely and connect with others on their own terms (Raman et al., 2019). According to Zignani et al. (2018), the utilization of a decentralized architecture allows for the examination of online social network dynamics and community formation without the influence of centralized platforms.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 671–676, 2025.
https://doi.org/10.1007/978-3-031-64431-3_112

The purpose of this study is to examine Italian-language content on Mastodon using standard textual analysis tools. In the beginning, we collected the 'toots' related to the hashtag #intelligenzaartificiale posted in the last two years, obtaining 340 posts after eliminating duplicates. The number of search results is limited due to several reasons. Firstly, the search is restricted to one hashtag and the Italian language exclusively. Secondly, searches must be conducted by selecting one instance at a time. In this case, we chose to search the instances mastodon.social and mastodon.bida.im. The purpose of our research is therefore not to obtain a complete representation of the Italian toots on artificial intelligence, but to analyze the quality of the material obtained through text mining and to experiment with machine learning procedures on a very limited set of texts.

2 Data Processing and Results

The R `vosonSML` library (Gertzel et al., 2024) was used to collect toots through the hashtag #intelligenzaartificiale for the years 2022 and 2023. Following basic text preprocessing operations, such as the removal of punctuation and numbers, a document-term matrix was created, which contained 3856 features. Furthermore, co-occurrence matrices were generated for hashtags and inter-user callbacks (exploiting @ symbol). The following graphs (Fig. 1) display the results of these two networks, highlighting the most recurrent connections. It is worth noting that (plot on the left) the hashtag network is particularly dense in the central area, with a few isolated nodes. Additionally, the bottom side of the graph contains references to OpenAI and ChatGPT, while the central body includes references to privacy, security, and data protection. However, when it comes to inter-user callbacks (plot on the right), these are very limited, with one clique in the right side and two cliques in the left side, connected by a node that acts as a bridge (@quinta).

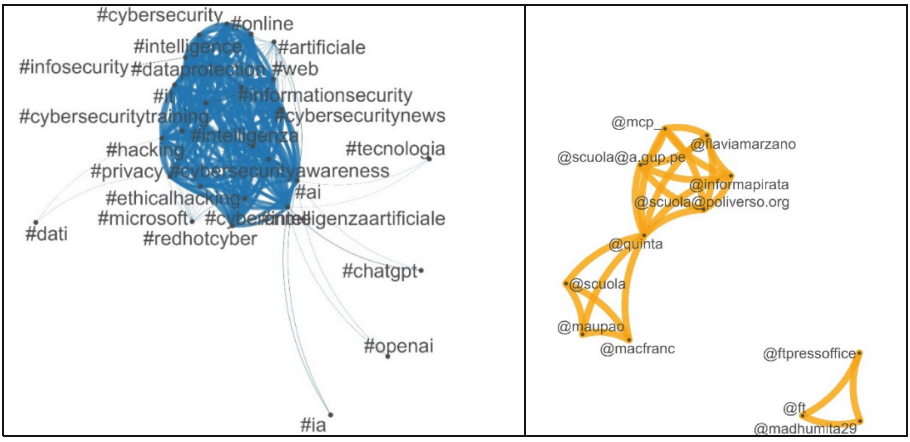


Fig. 1. Left: Hashtag co-occurrence network; Right: Users callbacks network

The following analysis concerns the topic modeling, which was performed through a structural model (STM, Roberts et al., 2016). At first, we removed Italian stopwords

and selected words that appeared at least 5 times. To determine the best number of topics, we used semantic coherence (Mimno et al., 2011) and exclusivity (Roberts et al., 2016) diagnostics, as combined through the Consistency and Differentiation (CD) score (Sciandra et al., 2023). The CD score calculates the L_2 norm of two min-max normalized measures to synthesize the trade-off between semantic coherence and exclusivity. According to Fig. 2, it appears that the ideal number of topics, as determined by the CD Score, is 4.

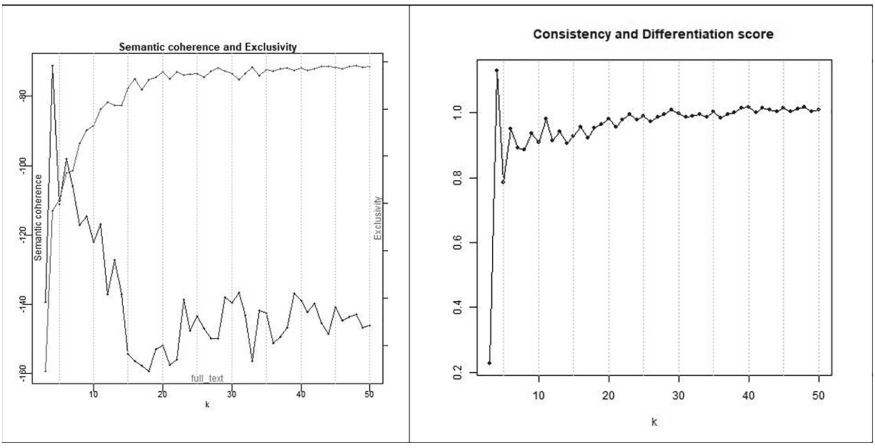


Fig. 2. CD Score for determining topic number.

The STM model includes a covariate indicating the year in which each toot was posted. This enables us to observe how the narrative of the topic of artificial intelligence may have changed, particularly since the diffusion of ChatGPT. According to the results depicted in Fig. 3, the first topic pertains to ChatGPT/OpenAI, while the remaining three relate to cybersecurity, privacy, and ethical hacking. By analyzing the most relevant words, topics 2, 3 and 4 seem to be quite related, which is confirmed by the existing correlations between them (Fig. 3 right).

During the final stage of the analysis, we tried to estimate the presence of engagement in Mastodon's toots related to artificial intelligence. The engagement variable is dichotomous and equals 1 if the toot received at least one reblog, favorite, or reply, and 0 otherwise. To estimate engagement, two textual data strategies were utilized. The first strategy involves utilizing word frequencies that reach a minimum threshold of 10 (69 words) as features. The second strategy involves utilizing a transformer to obtain a representation of the texts through a Large Language Model (LLM). In this case, the multilingual version of BERT (Devlin et al., 2018) was chosen, since the text contains many English words, hashtags, and Italian words. Since the BERT multilingual mapping contains 768 vectors, we applied a principal component analysis to reduce the dimensionality, extracting the first 69 components for a balanced comparison. For this purpose, XGBoost, one of the most commonly used machine learning models, was selected. XGBoost is a machine learning algorithm that is based on the gradient boosting algorithm developed by Chen et al. (2019). At a fundamental level, the algorithm



Fig. 3. Highest probability words for each topic and correlations between topics.

employs a sequential approach to improve the subsequent model via gradient descent. We highlight the regularization to avoid overfitting (through L_1 lasso regularization and L_2 ridge regularization) and tree pruning (as XGBoost builds the tree until the specified parameter, "max depth", starts pruning backwards) among the features that make the performance of this technique high. The training set was generated by randomly selecting 80% of the texts, while the remaining texts were used for the test set. Each model underwent 5-fold cross-validation repeated 30 times. The evaluation of the results included analysis of accuracy, number of misclassifications, and F1-score. Table 1 presents the results of the models for the test set. Both models perform well, although the model that utilizes BERT's vectors demonstrates slightly higher accuracy.

Black-box prediction algorithms (such as the XGBoost model used here) are criticized for their lack of explainability because they often provide accurate predictions without offering insights into how they arrive at those decisions, making it challenging for users to explain the reasoning behind the model's outputs. The increasing need for transparency and trust in AI systems drove the development of new methods, leading to the creation of techniques that provide insights into model predictions, facilitating understanding and interpretation by users and stakeholders. Among these, the Kernel SHAP (SHapley Additive exPlanations; Lundberg and Lee, 2017) is a method used in machine learning for explaining individual predictions. It leverages the Shapley value concept from cooperative game theory to assign each feature in a model a contribution to the prediction. Kernel SHAP computes Shapley values by estimating the expected value of the model output by considering all possible combinations of features. It does so by sampling subsets of features and evaluating the model's output, combining them with weighted averages. This approach provides a fair and consistent way to attribute predictions to features, offering insights into their importance and impact on model outcomes.

As an example, Fig. 4 illustrates the SHAP and explains how the prediction of (logit of the probability of) the engagement is composed by additive contributions of many features, in particular, by the presence of hashtags "redhotcyber", "online" and "privacy".

In conclusion, although the material collected by Mastodon was rather limited in size, text analysis and machine learning techniques revealed significant aspects of the discussion about artificial intelligence in this particular kind of social media. It is evident that this dataset cannot be considered a comprehensive representation of the Italian discourse about AI on Mastodon, given that only one hashtag was employed and that searches were conducted on only two instances.

Table 1. Results of the classification models for the engagement on the test set.

Features	Accuracy	N. of misclassifications	F1-measure (ref. ="Yes")
Words	0.8507	10	0.8958
BERT	0.8657	9	0.9126

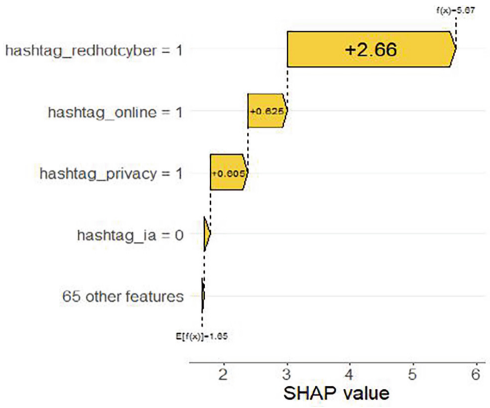


Fig. 4. Feature contributions for toot n. 10 (prediction: engagement)

References

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2019). xgboost: Extreme gradient boosting. R package version 0.81.0.1, 1–4

Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

Gertzel, B., Borquez, F., Graham, T., Ackland, R.: vosonSML: collecting social media data and generating networks for analysis. VOSON Lab, Virtual Observatory for the Study of Online Networks. R package version 0.34.1 (2024). https://github.com/vosonlab/vosonSML

He, J., Zia, H. B., Castro, I., Raman, A., Sastry, N., Tyson, G.: Flocking to mastodon: tracking the great twitter migration. In: Proceedings of the 2023 ACM on Internet Measurement Conference, pp. 111–123 (2023). https://doi.org/10.1145/3618257.3624819

La Cava, L., Greco, S., Tagarelli, A.: Information consumption and boundary spanning in decentralized online social networks: the case of mastodon users. Online Soc. Netw. Media **30**, 100220 (2022). https://doi.org/10.1016/j.osnem.2022.100220

Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst., 4765–4774 (2017)

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)

Raman, A., Joglekar, S., Cristofaro, E. D., Sastry, N., Tyson, G.: Challenges in the decentralised web: the mastodon case. In: Proceedings of the Internet Measurement Conference, pp. 217–229 (2019). https://doi.org/10.1145/3355369.3355572

Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. **111**(515), 988–1003 (2016). https://doi.org/10.1080/01621459.2016.1141684

Sciandra, A., Trevisani, M., Tuzzi, A.: Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment. In: Proceedings of the Statistics and Data Science Conference, pp. 61–66. Pavia University Press (2023)

Zignani, M., Gaito, S., Rossi, G.P.: Follow the "Mastodon": structure and evolution of a decentralized online social network. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12(1), Article 1 (2018). https://doi.org/10.1609/icwsm.v12i1.14988

Author Index

A
Agabiti, Nera 485
Agnoletto, Davide 3
Aguilera-Morillo, M. Carmen 611
Aiello, Luca 8, 665
Alaimo Di Loro, Pierfrancesco 14
Alfò, Marco 419, 561
Ambrosetti, Elena 208
Andreella, Angela 20, 27, 659
Arboretti, Rosa 46
Arcaio, Micaela 53
Arezzo, Maria Felice 58
Aria, Massimo 64
Armillotta, Mirko 71
Arnone, Eleonora 77, 170, 366, 426, 532, 611
Ascari, Roberto 82
Ascolani, Filippo 88, 93
Augugliaro, Luigi 254
Azzollini, Leo 98

B
Bacci, Silvia 104
Ballante, Elena 109, 115
Banzato, Erika 121
Barbi, Chiara 127
Bargagli, Anna Maria 485
Barone, Rosario 133
Barzizza, Elena 46, 139, 393
Bassi, Francesca 145
Battaglia, Salvatore 152
Battagliese, Diego 315
Battauz, Michela 158
Battista, Ilenia Di 532
Becarelli, Leonardo 164
Begu, Blerta 170
Bellanger, Lise 309
Bellani, Daniela 98
Bellomo, Michele 176
Belloni, Pietro 182
Benedetti, Elisa 189
Benelli, Giulia 640
Bernardi, Mauro 532
Bertarelli, Gaia 20
Bertolini, Marina 629
Bertolino, Francesco 260, 450
Biasetton, Nicolò 139
Biasetton, Nicoló 393
Biasetton, Nicolò 46
Bilancia, Massimo 196
Billari, Francesco 215
Bisaglia, Luisa 202
Bloise, Francesco 623
Bodo, Martina 208
Böhning, Dankmar 14
Bolano, Danilo 215, 412
Bollon, Jordy 221
Bombelli, Ilaria 226
Bondi, Laura 233
Bongiorno, Enea G. 239, 243
Borgoni, Riccardo 248
Borriero, Marco 254
Bortolato, Elena 260
Bove, Giuseppe 273
Brazauskas, Vytaras 279
Brombin, Chiara 285, 491
Brutti, Pierpaolo 386
Brutti, Zelda 291
Buccione, Roberto 491
Bufalo, Michele 297
Buonomo, Alessio 303
Burzacchi, Arianna 309
Buttigliero, Ylenia F. 93

C
Cafarelli, Barbara 555
Calculli, Crescenza 315
Caldura, Federico 20

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

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 677–680, 2025.
https://doi.org/10.1007/978-3-031-64431-3

Campanella, Francesco 480
Campostrini, Stefano 20, 27
Candila, Vincenzo 321
Canetti, Diana 176
Capecchi, Stefania 303
Capezza, Christian 328
Capizzi, Giovanna 328
Carbonara, Monica 334
Carella, Maria 341
Carey, Michelle 170
Carfora, Alfonso 347
Carluccio, Angela 315
Carone, Rossella 354
Casa, Alessandro 354
Casacci, Sara 334
Cascini, Silvia 485
Castagna, Antonella 176
Castiglione, Cristian 532
Castrovilli, Regina 360
Cavazzutti, Michele 366
Cavicchioli, Maddalena 373, 567
Cavinato, Lara 379
Ceccaroni, Riccardo 386
Ceccato, Riccardo 393
Cefalo, Leonardo 399
Centofanti, Fabio 328, 406
Cepni, Oguzhan 321
Cerrai, Sonia 189
Cesaroni, Giulia 485
Chan, Lax 239, 243
Chiavenna, Chiara 412
Chiogna, Monica 121
Ciardulli, Simone 354
Cipriani, Marta 419
Clemente, Aldo 426
Clementi, Letizia 611
Collesi, Daniela 431
Collesi, Patrizia 438
Colombi, Alessandro 444
Columbu, Silvia 450
Condino, Francesca 457
Conti, Pier Luigi 463
Corsini, Noemi 469
Costa, Michele 475
Costanzo, G. Damiana 513
Costanzo, Simona 561
Cotichini, Rodolfo 189
Cozzucoli, Lea Anna 34
Crisci, Anna 480
Cristofalo, Agostino 485
Cuccurullo, Corrado 64
Cugnata, Federica 491

D

d'Angella, Gabriele 519
D'Aniello, Luca 64
D'Elia, Silvia 561
D'Epifanio, Giulio 580
D'Onghia, Gianfranco 315
D'Orazio, Marcello 497
Dalla Chiara, Elena 497
Dalla Pria, Marco 503
Damato, Stefano 88
Dattilo, Barbara 513
Davenport, Samuel 525
Dawid, A. Philip 574
De Angelis, Roberta 592
De Giovanni, Livia 463
De Sanctis, Marco F. 532
De Santis, Fulvio 543
de Santis, Riccardo 525
De Stefano, Domenico 549
del Gobbo, Emiliano 555
Delbono, Flavio 475
Demaria, Fabio 567
Demuru, Giuseppe 574
Di Brisco, Agnese M. 82
Di Credico, Gioia 586
Di Iorio, Francesca 303
Di Mari, Fabrizio 592
Di Marzio, Marco 599
Di Nuzzo, Cinzia 605
Di Torrice, Mascia 513
Diotallevi, Sara 176
Disegna, Marta 393
Djordjilović, Vera 121
Dominici, Alice 617
Donelli, Pietro 611
Dorgali, Maria Veronica 617
Dotto, Francesco 623
Dreassi, Emanuela 104
Dudel, Christian 646
Dunson, David B. 3
Durante, Fabrizio 360
Duttilo, Pierdomenico 629

E

Elena, Grimaccia 266
Erba, Paola Anna 379

F
Fabbrucci Barbagli, Amin Gino 549
Fabrizi, Enrico 248
Failli, Dalila 635
Farcomeni, Alessio 133
Farnè, Matteo 640
Fensore, Stefania 599
Feraldi, Alessandro 646, 653
Ferraccioli, Federico 366, 659
Ferrara, Lorenzo 379
Figini, Silvia 109, 115
Finazzi, Francesco 665
Finos, Livio 366, 525, 538, 659, 671
Fiocco, Marta 152
Fontanella, Lara 555
Fop, Michael 469
Francescon, Nicola 354
Franzolini, Beatrice 233
Freddi, Matteo 354
Freo, Marzia 291
Frova, Luisa 653

G
Gall, Klervi Le 309
Galli, Laura 176
Gallo, Daniela 360
Gallo, Giampiero M. 321
Gardella, Jacopo 354
Gariazzo, Claudio 561
Garlanda, Elisa 354
Gattone, S. Antonio 40
Gattone, Stefano Antonio 629
Ghezal, Ahmed 373
Ghio, Daniela 215
Gialluisi, Alessandro 561
Giannantoni, Patrizia 303
Gianotti, Nicola 176
Gioia, Vincenzo 586
Giudici, Cristina 653
Giuli, Francesco 623
Goeman, Jelle J. 538
Goeman, Jelle 525
Goia, Aldo 239, 243
Gómez-Rubio, Virgilio 399
Gorgi, Paolo 71
Greselin, Francesca 279
Grilli, Leonardo 104
Guagnano, Giuseppina 58
Guandalini, Alessio 226
Gubbiotti, Stefania 543
Guetto, Raffaele 617
Guffanti, Monica 176
Gupta, Rangan 321

H
Hasson, Hamid 176
Hemerik, Jesse 525
Hennig, Christian 519
Hernández-Roig, Harold A. 611
Hubert, Mia 406

I
Iacoviello, Licia 561
Ieva, Francesca 379
Ignazzi, Elisa 555

K
Kume, Alfred 629

L
Latora, Salvatore 254
Legramanti, Sirio 8
Lepore, Antonio 328
Lillo, Rosa E. 611
Lolatto, Riccardo 176
Lombardi, Gabriele 189

M
Maiorano, Porzia 315
Marella, Daniela 273
Mariaelena, Bottazzi Schenone 266
Mariani, Francesco 543
Marinaccio, Alessandro 561
Marino, Maria Francesca 635
Martella, Francesca 635
Martìn, Juan Antonio Marmolejo 145
Massari, Stefania 561
Melegaro, Alessia 412
Menafoglio, Alessandra 127
Messina, Emanuela 176
Migliorati, Sonia 82
Mira, Antonietta 109, 115
Misuraca, Michelangelo 64, 513
Misuraca, Roberta 341
Molinaro, Sabrina 189
Mortarino, Cinzia 507
Muccini, Camilla 176
Musio, Monica 260, 450, 574

N
Nallo, Alessandro Di 215
Nigri, Andrea 196, 297
Nnanatu, Chibuzor Christopher 53
Nozza, Silvia 176

P
Paci, Lucia 8, 444, 665
Palma, Marco 233
Palumbo, Biagio 328, 406
Palummo, Alessandro 532, 611
Panzera, Agnese 599
Panzeri, Simone 170
Paparusso, Angela 208
Pappagallo, Marilena 653
Parroco, Anna Maria 53
Passamonti, Chiara 599
Paterno, Anna 341
Pauli, Francesco 34, 586
Pegoraro, Matteo 379
Perali, Federico 497
Pini, Alessia 354, 444
Pitrone, Antonio 438
Pizzolato, Federica 182
Putter, Hein 152, 538

R
Rafieisangari, Parvaneh 202
Rampichini, Carla 104
Rigon, Tommaso 3
Risso, Davide 121
Rivellini, Giulia 98
Rocci, Roberto 40, 592
Rocco, Lorenzo Di 419
Rossi, Silvia 592
Rousseeuw, Peter J. 406
Ruggiero, Matteo 88, 93, 503

S
Sacco, Giorgia 226
Sahu, Sujit K. 14
Salmaso, Luigi 139
Salvadori, Gianfausto 360
Salvati, Nicola 248
Sangalli, Laura M. 170, 366, 426, 532, 611
Sangalli, Laura M 77
Santelli, Francesco 549
Santis, Riccardo De 538
Sarnacchiaro, Pasquale 480
Scalese, Marco 189
Scandurra, Giuseppe 347
Scarlato, Margherita 623
Schirripa Spagnolo, Francesco 248
Sciandra, Andrea 671
Secchi, Piercesare 127
Serino, Luana 480
Serio, Clelia Di 285, 491
Serlenga, Laura 291
Spagnuolo, Vincenzo 176
Spanò, Dario 503
Spano, Maria 64, 513
Stamm, Aymeric 309

T
Tagliabue, Giovanna 592
Tiberi, Simone 221
Tocchioni, Valentina 617
Tomaselli, Venera 605
Tomasetto, Matteo 77
Torrisi, Gianpiero 605
Trapin, Luca 164
Trappolini, Eleonora 485

V
Valle, Alessandra Dalla 507
Vantini, Simone 309
Ventura, Laura 260
Vicard, Paola 285
Vidoni, Paolo 158
Vignoli, Daniele 617
Vinciotti, Veronica 254
Vitale, Domenico 58

Z
Zaccarin, Susanna 549
Zappa, Diego 248
Zemmouri, Imane 373
Zitikis, Ričardas 279